Colloquium on

SOLID STATE DEVICES

Sponsored by

Office for Industrial Associates
California Institute of Technology

February 20-21, 1961
Pasadena, California

Colloquium on

SOLID STATE DEVICES

Sponsored by
Office for Industrial Associates
California Institute of Technology

February 20-21, 1961
Pasadena, California

TABLE OF CONTENTS

"Microwave, Infrared, and Optical Masers," by George Birnbaum, Hughes Research Laboratories, is not included.

IMPACT OF SOLID STATE DEVICES ON THE ECONOMY

J. B. Fisk
Bell Telephone Laboratories, Inc.

At the turn of this century, science emerged as a potent force in man's affairs. The tempo of research was quickening and research was, with each passing decade, being more closely followed with applications through industrial engineering. This very close coupling of basic science and engineering has brought an increasingly powerful approach to man's understanding and to the control of his environment. By mid-century the teamwork between science and engineering appeared to be the greatest single force shaping mankind's affairs.

In the earliest times, people had little control of their lives and destinies. Mankind lay almost entirely at the mercy of nature and its forces. What little control people had obtained, they gained slowly through experience. They passed through periods of civilization - the Stone Age, the Bronze Age, the Ages of Steel and Steam - named characteristically by the tools and materials man learned to use.

In this sense, our present century must be known as the Age of Science. Far more than ever before, our tools and materials are the products of scientific thought and experiment.

Because tools and materials are the means of production - the very basis of any economy - the interrelationship of this entire scientific approach to economics is rather obvious. It is worth pointing out, however, that historically any tremendous new force such as this one, whether political, cultural or technical, eventually affects not simply one area of life, but the whole social fabric. Such a force is likely also to work great unexpected social changes on groups of individuals, particularly those who have shaped the force. I shall come back to this, but first let us examine the interrelations of scientific approach, electronics and the economy.

Fundamentally, research is concerned with new understanding of the world. Engineering is the application of this new knowledge to extend and improve our control of both the natural and economic environments. Today, scientists and engineers are both using the scientific method. This combination of scientific effort is, in fact, the major ingredient of the "growth industries" that are pacing our economic growth and sharpening our military strength. If we look about at the growth industries today, we do not see any which do not owe their position to a strong research and development program.

Most of our present "growth" products were in the research laboratory a decade or so ago. The outstanding example of this and the one we are concerned with here is the solid state production of the electronics industry. The fantastic growth of the electronics industry is based upon a large and firm foundation of basic research closely coupled through engineering to production and use.

In 1940 electronics ranked only fortieth among American industries. Today it is fifth or better. Also, it is still the fastest grower among all industries. Since electronics today has largely to do with the handling of information - high-speed acquisition, transmission and processing - it is a wonderful extension of man's mind. As such, this powerful tool is already permeating our whole economy and civilization through transportation, communications, manufacturing, the entertainment arts, medicine, business and warfare.

If we can measure the effect of the solid state sector of electronics in expanding the electronics industry, then by implication we can define the impact of solid state devices on the whole of our modern economy.

Let us review, then, the development of the electronics industry, first the epoch of the electron tube and then the solid state epoch. We shall attempt to subdivide these two epochs into shorter time periods according to the new economic applications that were made during these periods, which we shall call eras.

Prior to the audion's invention in 1907, communications was limited essentially to local voice-current telephony. By 1914, audion research was ready for large-scale development and production. The impetus was supplied during World War I in the accelerated applications of vacuum-tube amplifiers. These developments hastened the subsequent peacetime growth of long-distance telephony and the birth of radio broadcasting.

From 1917 to 1940, advances in tube power and frequency, coupled with the important concepts of modulation and feedback, brought the electronics industry to about the one billion dollar yearly level in 1940 - to about 40th place among all U.S. industries. The major application was radio and CHART I shows that this "Radio Era" was marked by steady but relatively slow growth. However, the apparent stability was misleading. Research potential had built up for further explosive growth under the impetus of World War II.

Under a massive coordinated attack by government, university and industrial laboratories, we developed high-power magnetrons, efficient klystrons, oscilloscope tubes and broadband grid tubes. These gave us long range and high precision radar over the complete spectrum up to 10,000 MCS. The same devices were put to work in many other military applications, but since radar was the dominant application, this wartime period from 1939 to 1946 has been dubbed the "Radar Era." Yearly production rose to a peak of three billion dollars. (CHART I)

At the war's end these developments went directly to work doing similar jobs in communications and navigation and initiated the rapidly growing electronic computer industry.

By 1950, these wartime advances had created a new electronic boom which, paced by television, rose to two and a half billion dollar yearly volume. The rapid growth of this "TV Era" was possible because of the very devices developed for military electronics.

From 1950 to 1956, the Korean War and the intensified "cold war" race for long-range, electronically equipped aircraft and for missiles marshalled in a new period of growth - the "Missile Era."

To process and transmit accurately the vast amount of data for global defense, high speed data processing and transmission systems were needed. By 1956 electron tube technology had been pushed to its practical limits.

All these data functions suffered from what has been called the "Tyranny of numbers" - because of their complex digital nature, the data functions required thousands of electron tubes.

The large amount of power (used inefficiently), the size and high cost of maintenance of tubes prevented the further rapid exploitation of these modern data transmission and processing concepts.

Indeed this is why the transistor promised such potentially important advances to electronics at the beginning of the solid state epoch, which I mark at 1956. It was obvious that transistors were very much smaller and consumed very much less power. It was also thought that semiconductor devices would not degrade as do the cathode or the vacuum of an electron tube. Because of their relatively simple mechanical features, they were thought to be potentially very low in cost - the other basic requirement in defeating the "tyranny of numbers."

Through intensive effort in many laboratories, great strides have been made in understanding the basic physics, in the purification and perfection of a number of materials, in the invention of new structures and in the development of new fabrication techniques such as diffusion and epitaxial growth.

The ranges of power and frequency performance have been extended to the point where more than 90 percent of the important applications of today and tomorrow can be handled.

Manufacturing techniques have been improved but not yet to the point where first cost is fully competitive to that of tubes - but rapid strides in mechanization and automation are being made so that it is easy to see the crossover coming in the near future.

In reliability, the struggle has been difficult, requiring much more basic understanding of the field of surface physics. It is fair to say that today the reliability of most semiconductor devices is several orders of magnitude better than that of even high quality tubes - even perhaps matching that of the expensive submarine-cable tube.

As a result, semiconductors are at work in a myriad of commercial, military and industrial systems - not just in the replacement of tubes in many of their traditional functions but particularly in those systems where tube limitations have prevented their economical application. In missiles, satellites, telephone transmission and switching, computers and all types of processing equipment, transistors are overcoming what appeared to be insurmountable barriers for electron-tube technology.

Semiconductor sales rose from essentially nothing in 1950 to half a billion dollars in 1960, greater than that of the competitive receiving tubes (CHART II). It is expected that by 1965 transistor-diode sales will reach a combined total of one billion dollars, which is more than the sales of all electron tube devices.

Semiconductor electronics is but the vigorous leader of the broad field of solid state electronics now developing. The very magnitude of this explosive economic growth of semiconductors has tended to overshadow the essential related evolution of other solid state components; yet they are vitally needed in association with transistors to make the newer electronic applications technically and economically feasible.

The large gains in semiconductor performance, size, cost, and reliability have forced our attention on associated passive components and methods for interconnecting them with semiconductors into functional electronic circuits.

Basically it is again the problem of numbers - too many components and too many connections - giving rise to large sized equipment, high cost, low reliability and limited performance.

To date the main attack has been through miniaturization, standardization and automation of assembly using such things as printed wiring boards and the micromodule method. However, this technique is not a basic attack on the problem of numbers.

Two other approaches are being made. The integrated circuit is based primarily upon the use of thin film techniques. It recognizes that it is not the number of elements that causes the trouble but the fact that they are handled as separate elements in manufacture and interconnection.

Large groups of high quality passive components and their interconnections can be made in one batch in a few operations. The chief weakness is that the active semiconductor devices must yet be appliqued as separately capsulated devices or as raw wafers, after which the whole circuit is singly enclosed.

Another approach, that of the functional, or molecular device, exploits our potential ability to perform electronic circuit functions by going directly to the physics of solids without being impeded by classical concepts of circuit elements. As our understanding grows, we can expect the invention and synthesis of single complex solid state devices which will replace in function whole circuit arrays.

Though not dignified by name, a few functional devices have been in use for many years, such as the piezoelectric crystal. Newer examples of functional devices are the PNPN diode and the PNPN array shift register or counter and the Esake tunnel diode.

With other electronic solids too, we have demonstrated the feasibility of performing complex logic and memory functions in a monolithic wafer of ferrite or ferroelectric material with large reductions in the component-connection count for a given function.

Beyond the goal of providing a better technology to perform electron-tube functions economically, solid state physics is now providing us with ferrite devices which simply have no counterparts in the electron-tube epoch of electronics.

Further, through understanding and application of the interaction of electromagnetic quanta and the discrete energy level electron spins of paramagnetic atoms in solids, the maser has achieved long-sought goals. It is an essential noise-free amplifier which makes space-satellite communications a practical reality over the large spectrum of microwaves and millimeter waves.

Lastly, but not finally, the application of these same maser principles has been pushed into the optical spectrum with the successful demonstration of a source of light waves potentially useful to broadband communications.

My description of solid state devices has not been complete either in quantity or in technical depth - nor should it be, for that is the role of my fellow speakers on this symposium. Rather it is intended by comparison and extension to the electron-tube technology to give you a feel for the tremendous upsurge of technical and economic capability that leads me to call the period from 1956 onward the "Solid State Epoch" of electronics.

To complete our evaluation of the economic impact of solid state electronics, what can we say about the technical and economic nature of its application in the Sixties - what kind of era will it be?

Our present age is characterized by its exponentially growing complexity - almost any measure demonstrates this.

There are more people and more machines. We need to communicate more often with one another and with our machines. Indeed our machines must talk to one another even more. As a result, population density, transportation speed, communication traffic, information processing volume and (we hope) our productivity are growing exponentially.

Man can only hope to cope with this increasing complexity - over large distances - through more efficient higher speed transmission and processing of information.

Solid state science promises us the technology to accomplish this end if intelligently applied. To borrow a term from the military, I shall call this communication-command-and-control application of electronics the "command function." It is my belief that this "Command Era" of electronics starting now will stretch indefinitely into the future - permeating every aspect of our lives - home, business, industry, and warfare. It will be needed for national welfare, just as the animal with the more advanced nervous system has a better chance of survival.

It seems likely to me, although industry forecasters do not always agree, that by 1970 the "Command Era" of electronics may be at least a twenty billion dollar industry based primarily upon solid state technology (CHART III).

Being human, we cannot help but speculate upon the more distant future, a pastime that is hazardous if not impossible. Just look at the past. Solid state electronics has been characterized by frequent and important research breakthroughs. Such new knowledge and inventions cause the subsequent technology to expand in a large step - large compared to the relatively slow growth between such steps. Thus, we can say that the solid state electronics industry has the following characteristics:

It is large and growing larger.
It is based firmly on science.
It is vigorous and highly competitive.
Its products are usable wherever information is processed, transmitted, or used - and this is virtually everywhere.

We have seen tangible, and I believe striking, evidence of the impact of solid state devices on the economy. We have touched on less tangible, qualitative relationships, and speculated a bit. Perhaps the matter should be left at that.

However, we can hardly discuss the "impact on the economy" without at least thinking of the interaction between, on one hand, the economy, along with the whole society which is closely related, and on the other hand, the new technology and the history and destiny of the people who have shaped it.

The roots of the technology and the histories of those who shaped it extend directly to the universities, which now as in the past, form the foundations of science. When we acknowledge our debts to science, we acknowledge our debts, at least in large measure, to such universities as the California Institute of Technology, which have produced the scientists. These are not short-term debts easily repaid. They are continuing obligations on us all - not simply material obligations, but loyalty to broader concepts. These include the traditional dedication of universities to search not merely for items of knowledge but for truth.

The obligations are particularly appropriate for us who find ourselves in the burgeoning electronics field. For, it may well be that these solid state electronics extensions to man's mind and to his means of control will yet have a greater impact upon society than the nuclear extensions to man's muscle; or perhaps greater than the exploration of space, which itself depends so heavily on solid state electronics.

If you will permit a personal observation, this prospect is truly humbling. It seems to me that a field of science and industry that has risen to such importance to our national economy and is still rising - and which also is so important to the survival of our whole way of life - such an industry has an obligation to acquit itself in the spirit of the new maturity it has found. In its new status, the industry inevitably must take its share of social and economic responsibility.

This extends also to individual members of the scientific and engineering professions. This century, as I mentioned, has indeed become the Age of Science. To a degree unknown in modern times scientists and engineers have been catapulted to a new status, gaining a long-sought respect alongside other professions.

With this new professional status go obligations, such as the ethical responsibility to avoid the kind of publicity that confuses concept with completion. Further, the forces that have been set in motion by the Age of Science are already influencing the lives and activities of those who are shaping it. Scientists and engineers are called on to provide leadership, not only in the areas of their technical competence, but in broader fields, in such areas as government and economics, which scientific technology has influenced so strongly. Scientists and engineers have won this new respect. With respect goes responsibility. Now they must shoulder the responsibility and use it - carefully and wisely.

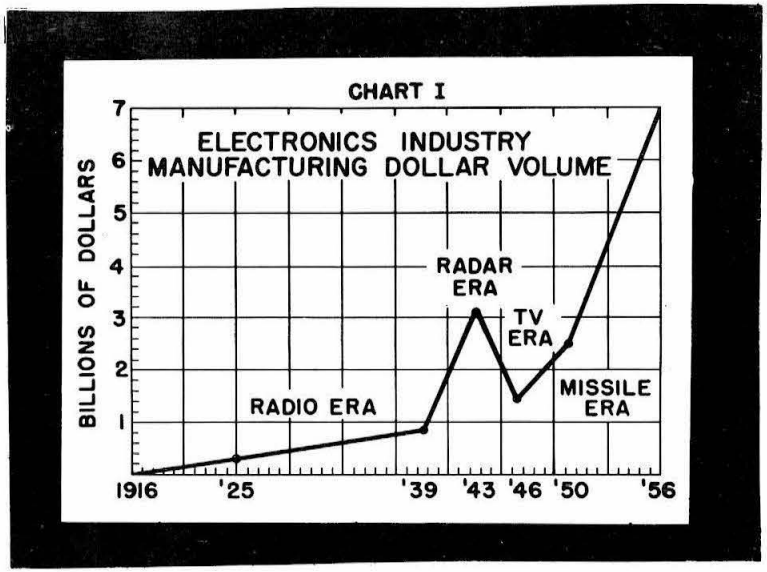


CHART I
ELECTRONICS INDUSTRY
MANUFACTURING DOLLAR VOLUME
BILLIONS OF DOLLARS
7
6
5
4
3
2
1
RADAR ERA
TV ERA
MISSILE ERA
RADIO ERA
1916
'25
'39
'43
'46
'50
'56

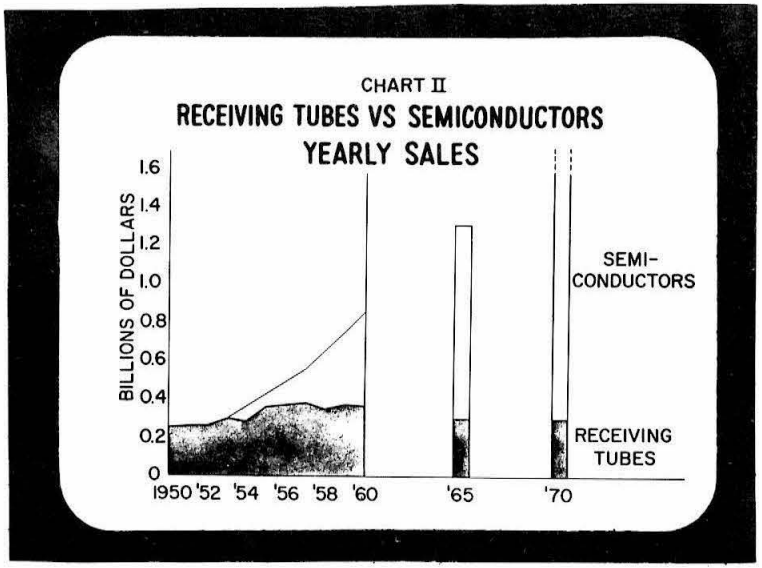


CHART II
RECEIVING TUBES VS SEMICONDUCTORS
YEARLY SALES
BILLIONS OF DOLLARS
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
SEMI-
CONDUCTORS
RECEIVING
TUBES
1950 '52 '54 '56 '58 '60
'65
'70

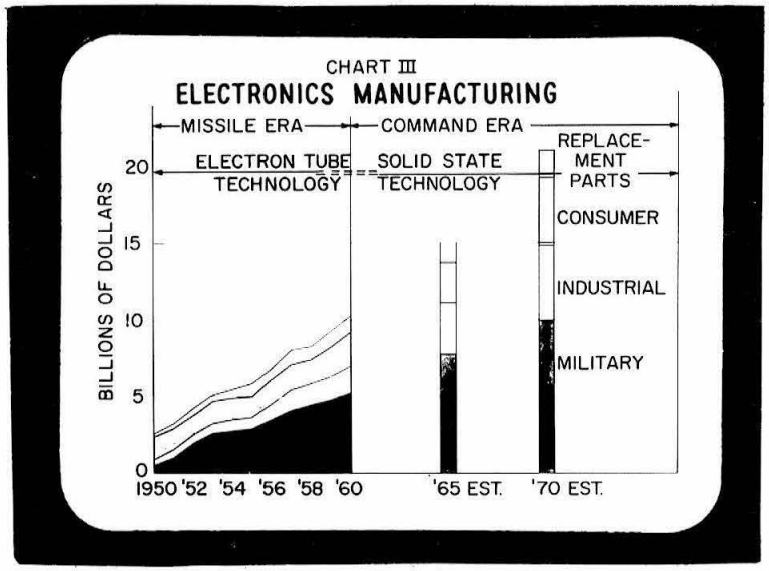


CHART III
ELECTRONICS MANUFACTURING
MISSILE ERA
COMMAND ERA
ELECTRON TUBE
TECHNOLOGY
SOLID STATE
TECHNOLOGY
REPLACE-
MENT
PARTS
CONSUMER
INDUSTRIAL
MILITARY
BILLIONS OF DOLLARS
20
15
10
5
0
1950 '52 '54 '56 '58 '60
'65 EST.
'70 EST.

SEMICONDUCTOR DEVICES

George C. Dacey
Bell Telephone Laboratories, Inc.

We would not have anything to talk about if we had not learned how to make fantastically pure and perfect crystals. Yet, it is the perfect crystal that is "bad" - at least not good; it is the departure from perfection on which we want to put our attention. However, I do not want to call it "imperfection." The French word is "étranger," which means "stranger," and this seems more appropriate. We shall call the departures from perfection in crystals strangers of one sort or another. So let us talk about some of the strangers that one finds in crystals. One of the strangers of importance that one might find takes the place of a germanium or silicon atom. This has a special name "donor" or "acceptor," which I think is familiar to most people. The conduction is by majority carriers, by friends rather than by strangers. They all belong together, and there is no way of knowing if an electron is put in at one point and looked at later whether the same one would come out at another point. Instead, the whole distribution moves. However, if one could paint one of these electrons and watch it move, he might be able to do something with that particular electron. With many conductors it is possible to do this because there is the other kind of carrier, the hole; it is a stranger or minority carrier, and it moves in the presence of its majority carrier opposite number (the electron friends) but in a different way. It can be "seen." It is a stranger and attention can be focused upon it. Transistors work because of the possibility of minority carrier flow.

There are electrons - a very "familiar" stranger in the world and those are not shown in this single crystal (Figure 1) except as those bars that hold the perfect crystal together. They are of no use in that form because they are not free to conduct. So it is the extra electrons that we want to talk about - friendly ones - but strangers to the perfect crystal. Extra electrons in a perfect crystal are strangers, and if there is one missing where one would normally be, that is a strange phenomenon too, and this is called a "hole."

Figure 2 is a schematic diagram which shows some of the "friends" and "strangers." For example, the large black circles are supposed to be the silicon or germanium, the semiconductor atoms. The small black ones that are tied down with the thin black lines to the large black circles are the electrons that are bound in the crystal. Shown is a free one and also the absence of one, the hole. The free electrons and the holes are strangers, and they are what makes crystals useful.

Let us talk briefly about electrons and holes and why this concept of having something different - and even having two different kinds - makes the semiconductor device game possible. Conduction in a wire is due to electrons. It is all electrons and they fill all the wire - at least in the simple theories. Hence, there are no strangers, so there are not any transistors out of wires.

Figure 3 shows a few more strangers; namely, those foreign atoms that I have discussed. This gives the way of controlling whether or not there are electrons or holes around in a controllable way. For example, suppose that the stranger we are talking about is the atom which has a little "5" in it. Germanium or silicon atoms from the fourth row of the periodic table have four valence electrons in their surroundings; they couple exactly to their neighbors. There is nothing strange about them, and thus all their black circles have 4's. However, there is one which has five electrons in its outer orbit; it cannot quite fit; it is a nonconformist. So its extra electron is free to wander around. These are called "donors" because they

donate their electron. If one puts in donors like arsenic or antimony which have five electrons, then for each one that is put in, there is one electron free to move around and conduct. Alternatively, if one puts in one of these circles with a "3"; say an atom of gallium or aluminum which has only three electrons in the outer orbit, it does not have quite enough to fit; it is also a nonconformist. It has to take up one from some other place and the electron it takes leaves behind a hole - another stranger - which is free to wander around through the crystal. In this way we have conduction by both electrons and holes, and as stated, this is what makes the semiconductor "device."

Figure 4 shows some more strangers. A perfect crystal would not have any electromagnetic photons in it. There was nothing in the previous diagram that said anything about that; and yet, suppose a photon does come in. It carries energy and can do things to the crystal. One of the things it can do is to go on through, and there are reasons why a semiconductor with things going on through is useful. It can be used for a window - infrared. On the other hand, with the right amount of energy, it can create a hole in an electron pair by knocking an electron out of where it should be. For a little while, the electron and the hole are free to run around in the crystal and conduct. Eventually, they will come back together again and recombine. When they do, the time they spent running around without recombining is called the "lifetime" of the crystal. The lifetime is very intimately related to the presence or absence of strangers, or friends, as one wants to call them, in this case; namely, atoms which help it to recombine. There are some other sets of names where the same concept as that of strangers comes in. They call them "deathnium" atoms because they make the charges recombine. Some call them "recombination centers," but the word that Shoppe used was deathnium. He could equally well have called them lifenium atoms because a diode or a semiconductor device that had an infinite lifetime in which the holes and electrons never recombined again would be no good and would not work. Suppose the diode were biased in the forward direction and some minority carriers were injected. If they never recombined and one tried to reverse the bias of the diode, the diode would never recover; it could not be shut off, and it would not be any good. So deathnium is not a good word because it means it might be bad. On the other hand, in some circumstances it is good; it is what makes the device work - the stranger atom that one puts in to cause that lifetime to go down.

Figure 3 also shows this situation where we have put in strange atoms in a controlled way. All the 3's are to the right of the line and none of the 5's, then all the 5's to the left of the line and none of the 3's. That makes a PN junction. The PN junction is the most fundamental device I know of in the general class called "semiconductor devices." The PN junction is important as everybody will agree. I am going to spend a little time talking about that one semiconductor device because although it is not called that in the rest of the program, many of the devices that are going to be discussed during the rest of the Solid State talks are, in fact, semiconductor devices. It is the kind of strangers that are in there that cause one to call it something else besides a semiconductor device. For example, if one is interested in the recombination of something, it might be called "infrared detector"; or if one is interested in how it behaves between the electrons and holes in tunneling, it might be called a tunneling device or an Asake diode, or it might be called an optoelectronic device. Many of these things which are sub-classes are also semiconductor devices; so in my semiconductor device talk, I will restrict myself to the ones which depend particularly on minority carriers, the injection of electrons and holes. In short, I will talk only about PN junctions with some holes and electrons on each side. We might first ask: Why don't the electrons diffuse around and fill up the whole crystal? If one pulls out the imaginary barrier that is put there, then the holes diffuse around and there is a uniform mixture of electrons and holes which neutralize each other's charge and which, again, is a useless device. The reason is that there are strange atoms in there which, once this process begins, prevent it from going very far. Suppose the

electrons from the N-side begin to diffuse over on to the P-side. As soon as very many have left, they leave behind - fixed in the atom - the opposite sign charge, a positive nuclear charge of the donor atoms from which they came. Conversely, when the holes leave the P-side and go over on to the N-side, they leave behind the acceptor atoms which are negatively charged. So at the boundary is a double layer of fixed charge - positive on the N-side and negative on the P-side, which prevents the further diffusion of electrons and holes. It is exactly that which makes the PN junction a rectifier because if there is no bias supplied to the junction, this diffusion process finally winds up with a barrier and no current flowing in the junction. If further bias is put on in such a direction as to try to push the electrons from the P-side over to the N-side, that same direction of bias, of course, will try to push holes from the N-side to the P-side; and one does not get anywhere because there are no electrons on the P-side; there are no holes on the N-side. Therefore, there is not much current, that is the direction of tough and hard current flow; it rectifies in that direction.

On the other hand, if the bias is put on the opposite direction so as to try to push electrons from the N-side where they are plentiful over to the P-side, then attempt to push holes from the P-side back to the N-side, this is much easier to do, and eventually enough voltage can be supplied to begin to overcome this double-layer charge at the boundary and the current will flow in an easy way. Therefore, the first thing that this simple set of strangers makes possible is essentially an ideal or perfect rectifier; at least, the most perfect rectifier that has ever existed - a PN junction.

Figure 5 shows a transistor, NPN, with two of these junctions. I think everybody knows how a transistor works. So, I will just say in the context of these strangers how the triode transistor works. If one of these junctions is biased in the forward direction so that one is pushing, in this case, across this junction in an easy way, and these electrons find themselves in the P-region, they are strangers; they should not be there. They are the minority carriers; attention can be focused on them, and it counts when they get across to J_ℓ. They diffuse across because there are no more junctions to stop them. Now, junction J_ℓ is biased in reverse and no electron comes back this way because this is the hard direction for this diode. However the electrons which came from this side are of the kind that I said you could not get when this junction was biased in reverse. They are electrons in a P-type region and so they go directly across.

So this is the way a transistor works: Electrons are injected across one junction where it is easy to do. Another way of saying that is: across a circuit in which the impedance is low. Let them go through a region where they live, where there are no recombination centers. In other words, keep them constant, let the current be continuous, and collect them across another junction which is biased in reverse where ordinarily it would be hard for current to flow - but now the electrons make the current flow all the same - but across a circuit where ordinarily it would be hard for a current to flow, and we can say has high impedance. Thus, there is a constant current that one can put in at low impedance and take out at high impedance so that power gain can be obtained. In a sense, that is all there is to a transistor.

What I have attempted to illustrate is that the understanding of the physics of semiconductor crystals in this instance has led to concepts of how these strangers interact, which makes possible this particular device and a whole host of other devices. Hence, this brings me to my closing line: Sometimes strangers can be friends.

DIAMOND STRUCTURE

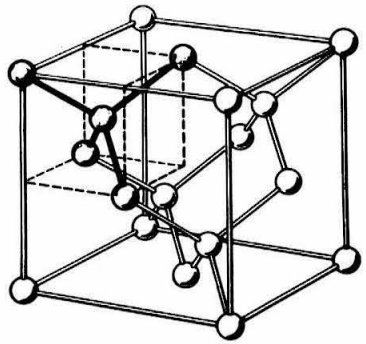


Figure 1

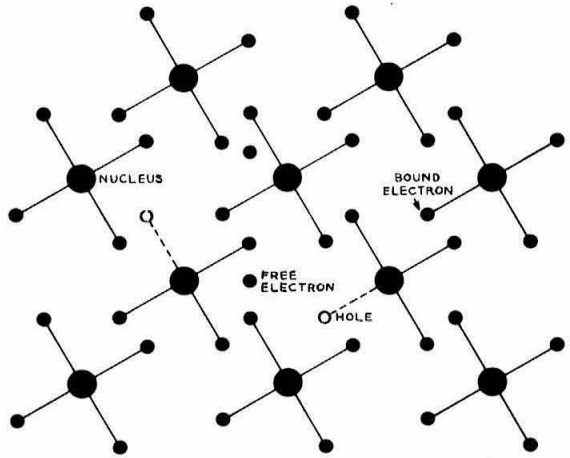


Figure 2

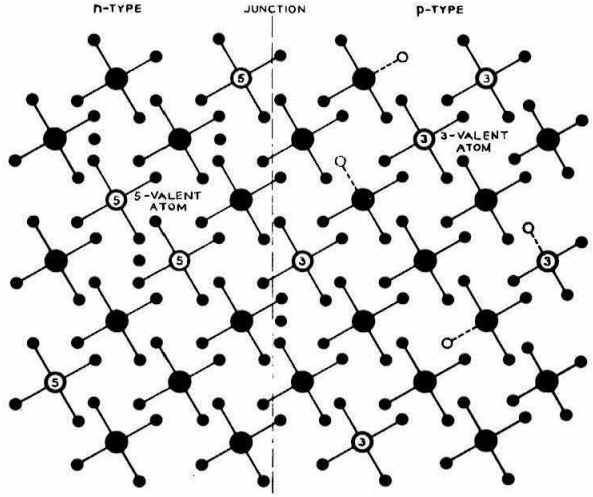


Figure 3

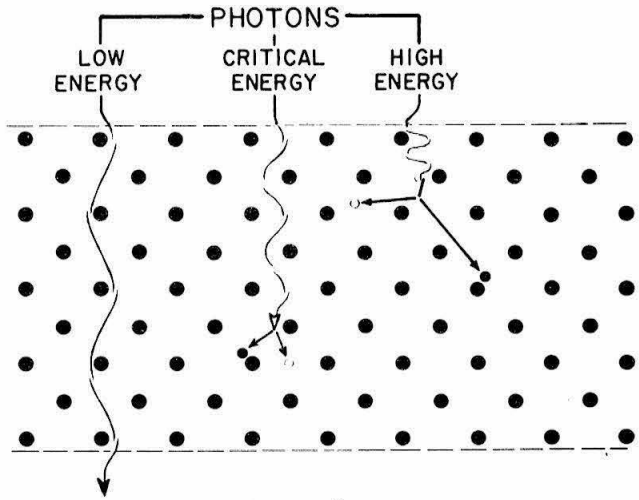


Figure 4

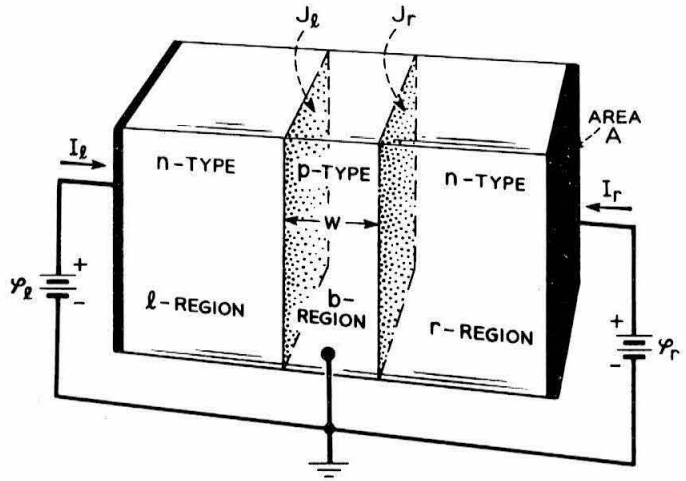


Figure 5

TUNNELING PHYSICS

Carver A. Mead
California Institute of Technology

I should like to discuss what I feel to be the two fundamental principles in our understanding of solid state physics to date. I think those of you who are experts on the subject will agree when I say that probably the outstanding features of solid state physics, as opposed to other branches of physics, are (1) that we deal with systems having a periodic structure of atoms or molecules of some sort - a crystal lattice, which is nearly perfect or nearly periodic over a large number of lattice spacings, and (2) that we deal with systems where the wave nature of the electron is of primary importance. I should like to put these two ideas together by considering what would happen, on a one-dimensional model, if we try to propagate an electron wave down through a periodic structure or lattice.

Let us first consider an electron nearly at rest. In this case, the wave length is very long and an electron can propagate unimpeded through a lattice.

Let us then give the electron more and more momentum - push it harder, down through the lattice. As the momentum increases, the wave length becomes shorter and we come to the point where the electron's wave length becomes equal to the spacing between the crystal's lattice points. We may assume that there is some type of interaction between the lattice and the electron wave. This can be called "reflection," or "scattering," or whatever one likes.

When the electron's wave length is equal to the spacing between the lattice points, this scattering, or reflection, from the first lattice point will be augmented by the scattering of the electrons from the second lattice point. A little bit more of it will be reflected from the next lattice point and a little bit more from the next. The reflections will all be in phase and the amplitude of the electron wave transmitted in the forward direction will approach zero. One can recognize immediately how this electron dies out going through the periodic structure of the lattice. If a certain fraction of the electron wave is scattered or reflected from each lattice point, this then is the amount of the decrement in the amplitude of the electron wave. When the decrement in the amplitude of the wave is proportional to the amplitude, the wave is damped out exponentially with distance. Our concept of the forbidden band in a crystal is basically that there are no propagating solutions for the electron wave with this certain momentum in this periodic structure. It does not mean that electrons with this energy cannot occur; it merely means that they will not propagate for any appreciable distance.

The tunneling physics that I would like to discuss is based on the behavior of electrons with this momentum. Figure 1 shows a generalized, solid state tunnel device. This device has two wires, one at each end with some material in between. This is a representation that could be a PN junction, where the material in between happens to be a crystal, half of which is P-type and half N-type. It could be only an insulator (which is what is shown) or it could be any one of a number of things. You can then ask: If there are electrons in the insulator, or a semiconductor which is in the middle, and if there are electrons on both ends, what sort of tunnel transitions can occur? The term "tunnel transition" as used here, means the transition in which the electron wave on passing (propagating) through a crystal (as from the left to the right in Figure 1), even though exponentially damped in the forbidden band of the crystal, is not completely damped because the distance through the forbidden band is very short. The electron wave is only completely damped if it must go an infinite distance

in the forbidden band. Therefore, if we make the "forbidden band" (the region through which it has to tunnel or through which it has to penetrate) short enough so that there is still some amplitude remaining at the other end, a certain fraction of the electrons manage to get through. There are many ways in which they can do this.

In a tunnel transition the electron wave can start in the metal, on the left of the diagram, and go through to the conduction band of the insulator or the semiconductor (3). They can start in the valence band of the semiconductor and move into the conduction band of the semiconductor (1) or they can start in the valence band and move into the other metal (4). All these mechanisms are possible wherever there is an electron which has only a finite distance to go until it reaches a position where it is allowed to have a propagating solution. These are so-called "tunnel transitions" - or electrons penetrating through the "forbidden region."

An Esaki tunnel diode is a device which looks very much like the energy diagram in Figure 1, except that there are some kinks in it. Basically, we are looking at transitions from the valence band of the semiconductor into the conduction band. One experiment, which I should like to tell you about has been done at numerous laboratories around the country and is quite illuminating to our understanding of tunnel physics.

Consider a device which has a forbidden band through which electrons are tunneling. In reality, we never have a one-dimensional type of periodic structure. We have a crystal which works in three dimensions; therefore, our placement of the conduction and valence bands, or the energies at which the electron has a wave length which just allows it to be totally reflected, vary in the different directions. Consequently, if we talk about valence and conduction bands, for example, it is not necessarily true that the valence electron with the highest energy is going in the same direction as a conduction electron with the lowest energy. If one wishes to make a transition in a semiconductor device which has an electric field in it where the electron has to penetrate the minimum distance, this might be a transition where the electron has to be deflected - where it is traveling in a different direction when it is through making the transition than when it started. If this is true, the electron must somehow give up momentum or be given momentum. Of course, at room temperature we have plenty of "strangers" in the lattice called "lattice vibrations" or "phonons," which are just quantized lattice vibrations, so that the electron has no trouble doing this. If we cool the diode down to say, 4° Kelvin, or somewhat less, we soon find that it is very difficult for electrons to do this. The number of electrons which are making tunneling transitions - from going in one direction in the valence band to going in another direction in the conduction band - is very small. If we apply a certain amount of voltage, there is a very small number of electrons which are allowed to do this - until we reach a certain critical voltage. This voltage corresponds to the energy required to create one of these little quantized lattice vibrations. At this point the electron has enough extra energy to enable it to excite one of these quantized lattice vibrations in order to change its direction of motion. Consequently, it is much easier for it to make this tunneling transition (and the current suddenly increases), as shown in Figure 2.

We can then go a little further and nothing much will happen until we reach the point where another one of these quantized lattice vibrations can be excited and the current again increases very rapidly. This study of tunneling and its relationship to quantized lattice vibrations has given us a better understanding of the tunneling phenomenon itself and also of the basic nature of the "strangers" within the semiconductors that we are talking about.

For the rest of this talk, I should like to concentrate on materials which are not semiconductors. I shall cite two experiments which I consider to be of some interest. The first of these experiments was done at the General Electric Company, and the second was performed here at the Institute.

Figure 3 shows the energy diagram for a structure which consists of a metal, a very thin insulating layer, and another metal. Can we obtain tunneling transitions from one metal to the other? Notice in this diagram that the number of electrons which can make tunneling transitions is only the number of electrons in the eV shown. The reason for this is that the electrons in the metal on the left, which are at energies below the Fermi level on the right, are all facing filled states and the exclusion principle does not allow them to make a transition. So it is only electrons above this energy which can tunnel into vacant states shown on the right. Since the number of electrons facing vacant states on the right is roughly proportional to the voltage applied, we would expect the volt ampere characteristic of this kind of device to be ohmic (obeys Ohm's law).

There is an exception to this. If we have a material which has no vacant states on the right-hand side - if by some mechanism we are not allowed to have electrons for a certain range of energies above the Fermi level - we will then have to apply higher voltage to get out of the region where there are no allowed states. This situation exists if, on the right-hand side, we have a superconductor. It is known nowadays, from some recent work done by Bardeen, Cooper and Shreifer, that a superconductor should have a very small energy gap or a very small increment of energy in which there are no allowed electron states just above the Fermi level.

Figure 4 shows what happens in such a device. At room temperature or so, it has ah ohmic characteristic as expected. However, if we make a superconductor out of it, you see that for a small number of millivolts - corresponding to the width of this forbidden gap on the right-hand side - we are not allowed to have electrons traveling over to the right because there are no vacant states into which these electrons can tunnel; therefore, there are no electrons making transitions and there is no current. After the applied voltage exceeds the forbidden gap, the current increases very rapidly because the electrons then have allowed states into which they can tunnel. In this way, it is possible to observe directly the energy gap in a superconductor. There have been some extensions of this experiment using two superconductors, and it turns out that a negative resistance can be observed in this way.

Figure 5 shows the conductance of this same device as a function of voltage. This conductance is simply the differential of the voltage current characteristic. It gives essentially the density of the vacant states that are allowed for electrons in the right-hand metal. There are none for some time, then the density of states goes up very rapidly as if, in a superconductor, the states that normally exist all the way down to the Fermi level are somehow crowded out just above the forbidden gap. The solid curve was computed from the Bardeen, Cooper-Shreifer theory. It exhibits almost too close agreement with the experimental points.

It will be noticed in Figure 4 that the distance through the forbidden gap through which the electrons had to tunnel was essentially constant. This being true, the number of the electrons left over after getting through the forbidden gap was essentially constant and the nonlinearities in the voltage current characteristic came from the difference in the density of states. Some work has been done here at the Institute in which we have looked at the other extreme. We have made the insulator relatively thick so that a sizeable electric field has been applied to get the path through which the electrons have to tunnel down to where a substantial number of electrons are tunneling. The electrons make the transition, labeled 3 in Figure 1, from the vicinity of the Fermi level in the left-hand metal to the conduction band in the semiconductor, or insulator.

Figure 6 shows this kind of experiment in a little more detail. We have a metal on the left, a metal on the right, an insulator in between, and an electron wave schematically representing the kind of transition taking place. An electron travels to the right with a large amplitude, is exponentially damped in the forbidden region, then gets into the conduction band of the insulator gathering energy as you can see when the wave lengths get short, and finally entering the second metal.

Now, there are two questions we should like to ask about this experiment. What will the voltage-current characteristic look like? What happens to the electron when it gets over into the second metal? The first question is quite easy. The amplitude of the electron wave, (the fraction of the electrons that get through) is exponentially dependent upon the inverse of the distance that it has to go. The distance is just the work function, (the height above the Fermi level where the semiconductor condition band begins) divided by the electric field. So one would expect the current, which is proportional to how many electrons make this transition, to go like 1/electric field.

Figure 7 shows some experimental points with a fit to a theoretical curve. It turns out that the people who do a lot of theoretical work have difficulty getting within factors of 10^4 of the kind of current actually observed, so a constant, one way or the other, does not seem to bother anybody. I took the liberty of adjusting this curve up and down to make it fit the data. As you see, there is a reasonable fit between the data and the experimental points from this kind of device.

The temperature dependence is shown in Figure 8. If one takes the voltage required for a given tunneled current as a function of a temperature, it is found that it is constant at low temperatures and then falls off at 1/T at higher temperatures.

Another question is, of course, what sort of energy distribution do these electrons which make this tunneling transition have? That is shown in Figure 9.

Above the Fermi level in the left-hand metal we rapidly run out of electrons to make this transition, so the current goes to zero (this was actually done for absolute zero temperature); at higher temperatures, there is a little "tail" up there. At lower energies the distance through which the electrons have to tunnel is larger, so the current rapidly decreases. Hence, we have a source of electrons making a tunnel transition which are reasonably monoenergetic and whose current density is controlled by the voltage that applies to the diode.

I would like to come back to the question I asked before, and that is: If you have a metal-insulator-metal structure, what happens to the electron after it has gone through the insulator and it is over in the second metal? It has a lot of energy, corresponding to the voltage applied to the diode. Eventually one knows that the electron wanders around in the metal and ends up in equilibrium down by the Fermi level somewhere. But how long does this take? How far does the electron have to go before it has collisions which decrease its energy? I have asked a number of people and received the same number of answers. Some said that it went a very short distance - just a few atomic diameters. Others said it went a large distance; and the rest said that it was somewhere in between. It turns out that all answers are right, as you can see in Figure 10.

Figure 10 is a plot of some work done by Harry Thomas in Germany a few years back with potassium. We have good evidence that the same thing happens in other metals. Plotted is the mean free path (not the conductivity mean free path, but the distance the electron has to go into this metal before it actually has a collision and loses energy) as a function of electron energy. This mean free path is a very sensitive function of just how fast the electron is going. The electrons with low energies have very long mean free paths. It can be seen that they are on the order of 1000 angstroms;

then as the energy is increased toward the plasma resonance energy of the metal, the electron mean free path decreases very suddenly because the electron is exciting plasma oscillations in the metal. Then we ask ourselves the question: What would happen if we made that second metal thin? - compared to a thousand angstroms - which is not hard to do. Would the electron go right on through and come out the other side? The answer is, yes it would. The energy band structure of an experiment like this is shown in Figure 12.

Shown to the left in Figure 11 is the metal-insulator-metal structure. In the middle labeled "Base" is a metal layer that we have made very thin compared with the mean free path of the metal at that energy. Between the Base and the Collector we have put another insulator or a vacuum. On the right is something to collect the electrons which come out through the metal film (Base). It is hard to do these things, and Figure 12 shows how we go about building them. We start with the metal (Emitter) and cover up the edges so that there is just a little spot in the middle where we want to do the experiment. We anodize a very thin insulator on top of the metal and evaporate on it a thin metal layer (Base). Then we either add another insulator, or if we do not want another insulator, we put it in a vacuum system with a metal plate facing the structure and apply an electric field to that. It turns out that if we actually build these devices, they work. They do not work well yet, of course, because we have not built enough of them and we do not yet know very much about them. However, one of the particular characteristics we have observed (and we think we are learning a little of solid state physics out of this) is shown in Figure 13.

In figure 13 is plotted the fraction of the electrons which actually tunnel between the first and second metal layers - the fraction that comes out into the vacuum or into the second insulator layer and is collected at the other end as a function of the total emitted current - the current through the first metal layer. Universally, in any of these devices we have tested, the fraction gets bigger as the current gets bigger. We think this is because there are a lot of interface states and many ways for the electron to get trapped and not be able to get through. We cannot yet make these devices out of nice single crystals because it is a very difficult technique. If we force more and more current through the device, the electrons have a better chance of overcoming these trapped states and of getting through to the other side.

By the use of this technique we hope to learn something about the properties of solids, and in particular, about the behavior of electrons of a few electron-volts energy in metals and insulators. Also one would hope that practical application for these principles could be found. To date, there has been a great deal of speculation but it would not be surprising if cold cathodes of very high current density and high frequency triode amplifiers using this principle could be developed. However, the materials and technology problems involved are very difficult and a large effort will undoubtedly be necessary before practical devices become feasible.

Figure 1

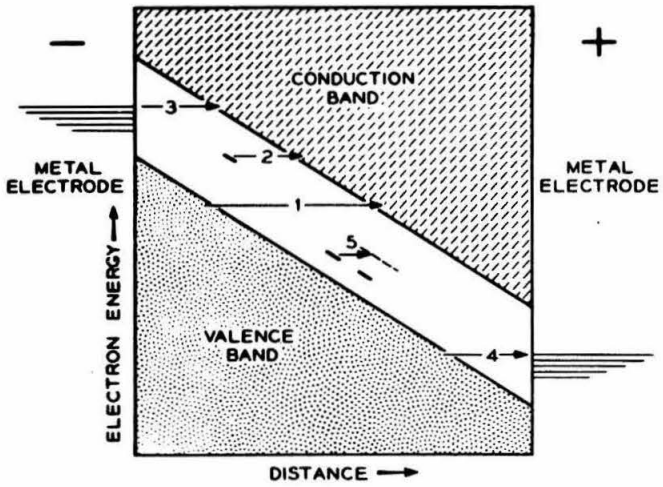
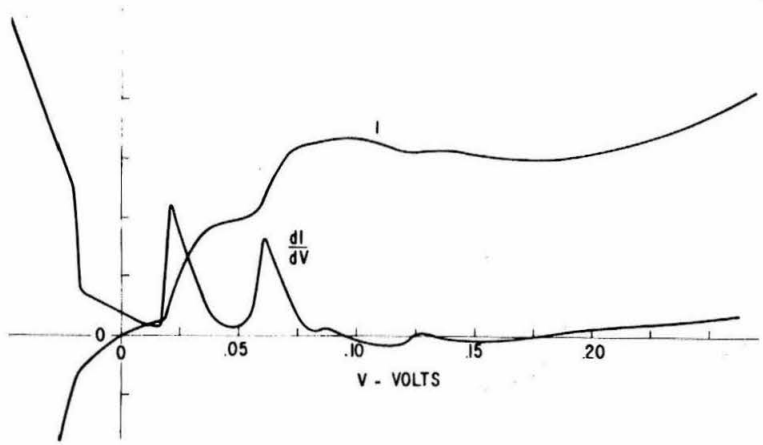


Figure 2



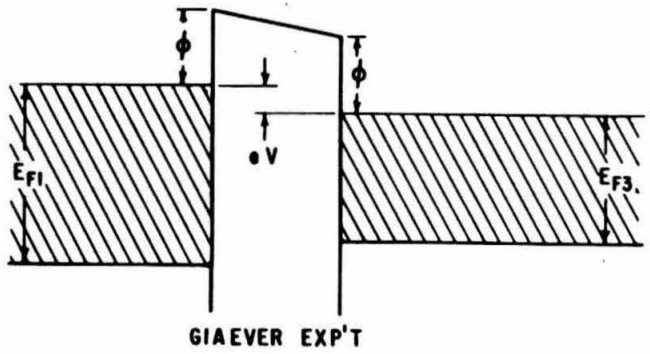


Figure 3

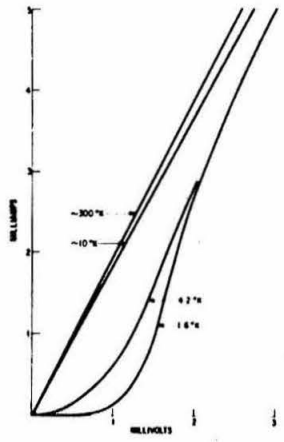


Figure 4

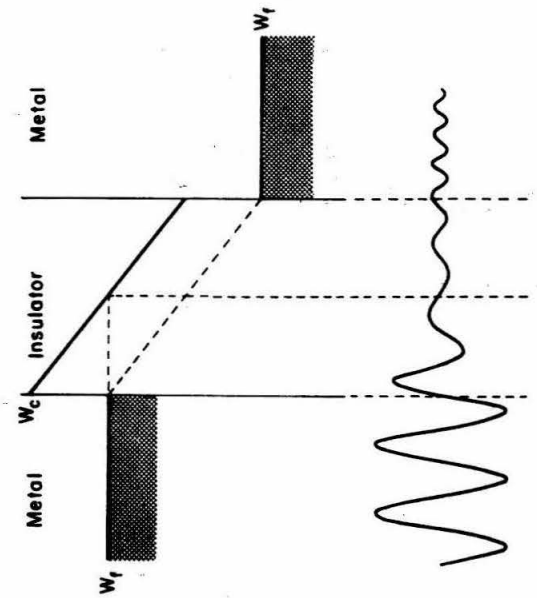


Figure 6

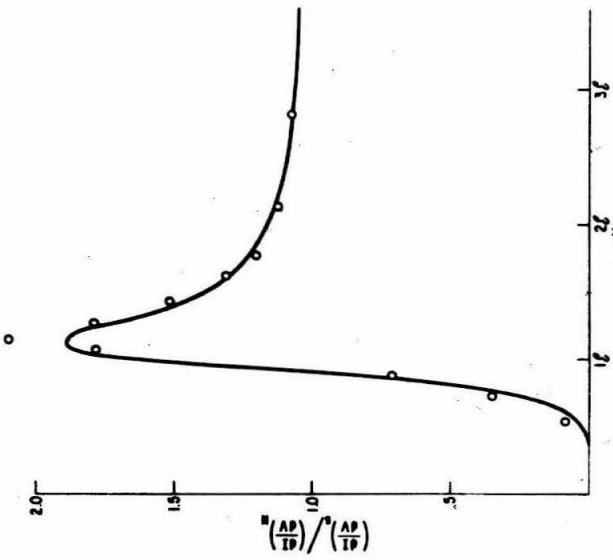


Figure 5

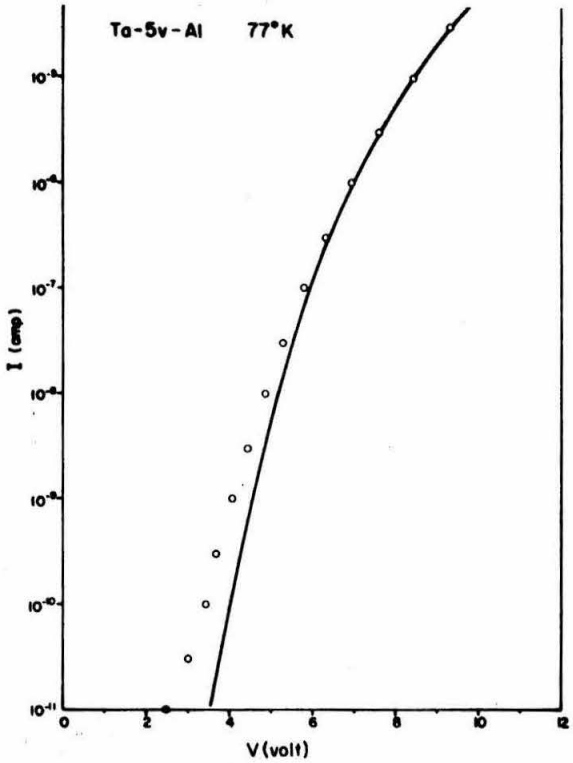
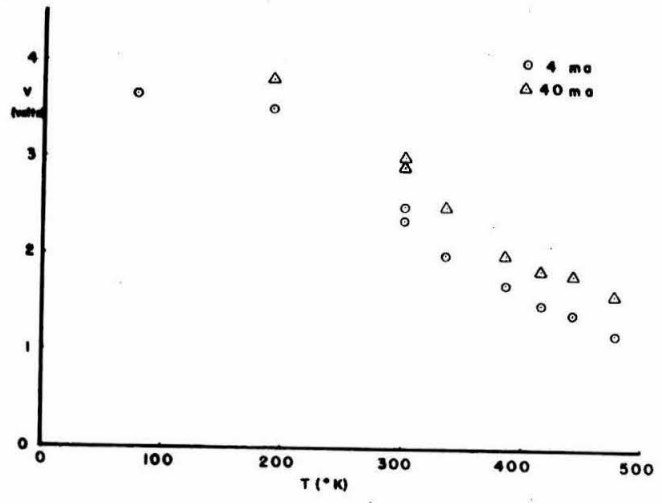


Figure 7

Figure 8



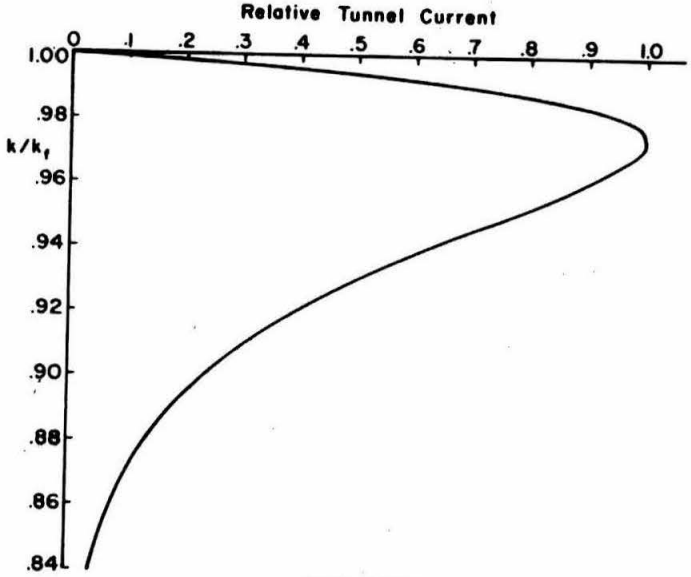


Figure 9

Figure 10

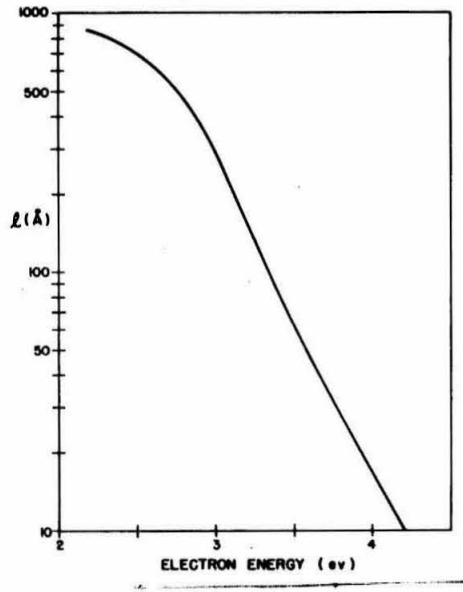
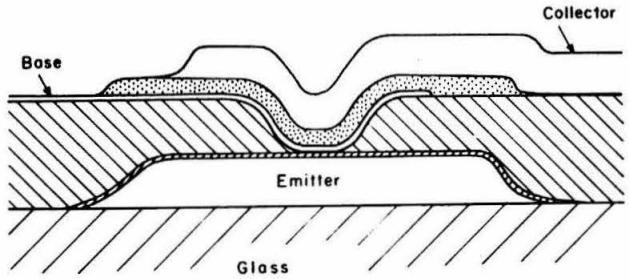


Figure 11



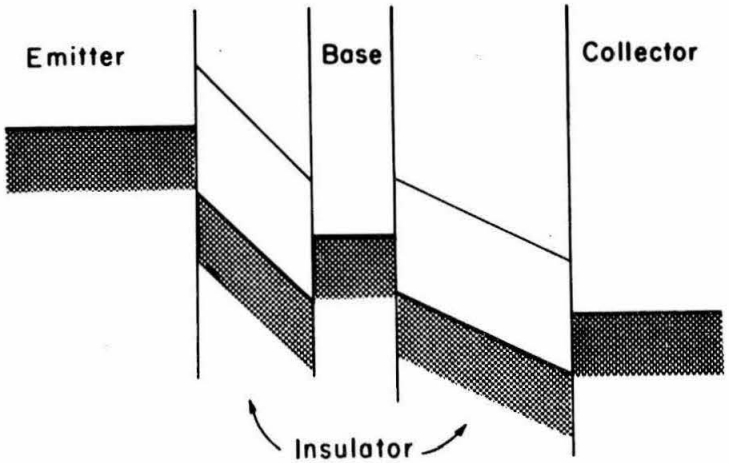


Figure 12

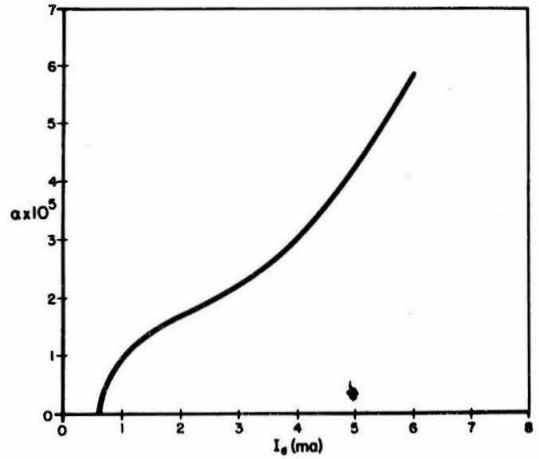


Figure 13

FERROMAGNETIC DEVICES

Floyd B. Humphrey
Jet Propulsion Laboratory

Ferromagnetism is certainly not new in the electronics field. Ferromagnetic devices in the form of inductors, transformers, and relays were used in early electronics. Therefore, these kinds of devices and the kinds of investigations that lead to an understanding of these devices will be defined as old-hat and will not be discussed. Instead, the new applications of ferromagnetic materials, uses and applications which, for a large part, grew out of the digital computer industry, shall be concentrated upon. Magnetic devices such as cores are very popular as computer memory elements even though they are relatively new. It has been only ten years since Forrester[1] suggested their use as a random access matrix memory. Even then the cores were metallic, i.e., tape-wound cores; it has only been eight years since J. A. Rajchman[2] demonstrated that a reasonably good size (10 thousand bits) ferrite core memory could be made. This restriction to computer magnetics leaves out the microwave applications of ferromagnetic material even though many applications are new and exciting. The justification for this action is that the subject will be adequately covered in a subsequent symposium.

For operation in a digital computer, a device, typically, has two stable equilibrium states. During the operation, the device is switched from one stable configuration to the other. Such devices have been the motivation to investigate and to understand in detail the mechanism of the rapid magnetic flux reversal. Menyuk and Goodenough[3] formulated a domain wall motion model of flux reversal for ferrite and metallic tape memory cores. Gyorgy[4] recognized that such a model was good only for low drives and proposed a nonuniform rotation model for the flux reversal in cores. It was then recognized by Humphrey and Gyorgy[5,6] that the reversal was even more complicated, requiring three mechanisms depending upon the magnitude of the reversing field (drive), and that such a model was applicable to all soft ferromagnetics. It is this current model of flux reversal which will be discussed in more detail.

Since this model is in general, applicable to all soft ferromagnets, the liberty will be taken of discussing mostly thin ferromagnetic films. These films are typically 2000 A thick of 83% nickel and 17% iron evaporated onto a hot (300°C) glass substrate. If there is a magnetic field in the plane of the film when the film is made, there will be a uniaxial magnetic anisotropy with the easy axis in the direction of the field. The origin of the anisotropy is not understood. It greatly influences the magnetic characteristics of the film. As is indicated schematically in Figure 1, the direction parallel to the easy axis of the anisotropy is called the longitudinal direction and the direction perpendicular is called the transverse direction. When there are no external fields, the magnetization will be parallel to the easy axis, that is, magnetized in the longitudinal direction.

A typical flux reversal experiment consists of setting the film in the longitudinal direction with a field many times the coercive force and then removing the field so that the film will remain in one of the two remnant states fully magnetized along the

---

1. Forrester, J. W., Jour. Appl. Phys. 22 44 (1951)
2. Rajchman, J. A., Proc. I.R.E. 41 1407 (1953)
3. Menyuk, N., and Goodenough, J. B., Jour. Appl. Phys. 26 8 (1955)
4. Gyorgy, E. M., Jour. Appl. Phys. 28 1011 (1957)
5. Humphrey, F. B., Jour. Appl. Phys. 28 284 (1958)
6. Humphrey, F. B., and Gyorgy, E. M., Jour. Appl. Phys. 30 935 (1959)

easy direction. Now if it is plunged instantaneously, say in a few tenths of a millimicrosecond, into a uniform field of magnitude H, which is in the direction opposite to the setting fields, the magnetization will switch to a new direction. Loops can be placed around the film to infer the flux change by observing the induced voltage.

The time τ will be defined as the time to switch from 10 to 90 percent of the integrated flux. Figure 2 illustrates the results of a typical experiment. Here one over the reversal time τ has been plotted against the drive field H for various transverse fields H_t. The three regions are indicated by the curved lines, the solid lines and the dashed lines. It takes three different mechanisms to describe this behavior, one mechanism for each region.

In the first region, the low drive region, the mechanism of flux reversal is thought to be by domain wall motion much the way Menyuk and Goodenough described. When a field H over a certain threshold is applied to the sample, domain walls sweep through the sample as one domain grows at the expense of another. The threshold can be recognized as being related to the coercive force. Over this threshold the velocity of the walls increases as the drive increases, reducing the switching time. The most striking indirect evidence for wall motion is the interrupted pulse experiment first done on toroids by Gyorgy and Rogers[7] and repeated on films by Hagedorn[8]. With the drive adjusted such that the reversal is in the domain wall motion region, the field pulse was interrupted in the middle of a reversal. The model predicts that the walls will stop and then start again when the field is reapplied and move as if there had not been an interruption. Figure 3 illustrates the results of this experiment. This is the data of Hagedorn. The upper trace is a typical voltage transient observed during a flux reversal. If the field is interrupted as in the lower trace and then restarted, it can be seen that the induced voltage follows on as if nothing has happened.

The most convincing direct evidence for domain wall motion is to see the domain walls move. Thin films are particularly adaptable to such an experiment since they can be used, without any surface preparation, to see domains using the Kerr magneto-optical technique. As can be seen in Figure 4, the direction of polarization of incident polarized light is rotated depending upon the direction of magnetization of the films. Typical domain patterns can be seen in Figure 5.

Referring to Figure 2, there is a second region that is always linear, called the nonuniform rotation region. Here the interrupted pulse experiment does not work, as can be seen in Figure 6. It takes as much time to complete the reversal after the interruption in B as it did for the whole sample in A. Such evidence clearly suggests a mechanism different from domain wall motion. The nonuniform rotation model proposes flux reversal in the intermediate drive region by a rotation process where the phase of the precession is not preserved throughout the sample but changes from one spot to another in the material. This breakup is similar to magnetization modes or spin waves. Their presence allows local cancellation of the demagnetization field resulting in a lower energy path; lower, that is, than the uniform rotation. The model predicts that the shape of the switching transient, that is, the observed voltage ($V \approx \dot{M}$) will be proportional to the hyperbolic secant squared. The fit is quite good as can be seen in Figure 7, which is the case for films. The model also predicts that the maximum slope of the $1/\tau$ vs H curve (with zero transverse field) will be dependent on the g-factor of the material, and for most ferromagnetics where g = 2, the maximum slope = 5. Such seems to be the case, since films (with zero transverse

---

7. Gyorgy, E. M., and Rogers, J. L., Proc. Mag. Conf., Boston (1959) AIEE T-91, p. 637
8. Hagedorn, F. B., Jour. Appl. Phys. 30 254S (1959)

field), ferrites, garnets, and tape cores all have slopes close to but less than 5.

In Region III, the high drive region, the mechanism is uniform rotation. Here a single vector can be used to represent the entire magnetization of the sample. The motion of this magnetization vector in the presence of an applied field is described by the Landan Lifshitz equation. The motion of the magnetization is very much like the motion of a gyroscope consisting of an (M x H) term plus a damping term. In a thin film, then, we would expect the magnetization to precess up out of the plane of the film and then just swing around, damped only by the intrinsic damping of the material. Such a process should allow very fast flux reversals. For loops arranged as in Figure 8, the voltage induced in the transverse loop should change sign when the film is half-switched in the longitudinal direction. The longitudinal and transverse voltage transients have been observed by Olson and Pohm[9] and also by Dietrich, Proebster, and Wolf[10]. The latter data is reproduced in Figure 9. The evidence supporting the uniform rotational mode is generally not this good. The fault is only partly the model since most of the data is taken in times so short that the experimental problems are formidable. The model gives a reasonable explanation for the existence of the short reversal times observed, although it generally predicts times shorter than those measured, and detailed agreement between the theoretical and observed switching wave forms in the longitudinal direction is very poor.

In summary, then, it has been shown that at least three mechanisms must be considered in describing flux reversal in soft ferromagnets. In Region I, where the drive field is only slightly larger than the coercive force, flux reversal takes place by domain wall motion. The most convincing evidence is the direct observation of the domain or domain walls and the interrupted pulse experiment. Larger drive fields lead to Region II, where most of the experimental observations are consistent with a nonuniform rotation model. In particular, such a model predicts the linear relation between drive field and the inverse of the reversal time, the observed shape of the flux reversal transient, and the observed value of the maximum slope. Uniform rotation provides the best description of the observations of flux reversal in Region III, where a still larger drive field and a transverse magnetic field are required. The observed fast reversal is accounted for as is the transverse switching transient.

It is reasonable to ask about the devices whose behavior is described by our switching model. Although the domain wall motion model was first proposed to describe the operation of memory toroids, such as the thin film shift register that was proposed by Moore[11] and demonstrated by Broadbent and McClung[12], it more clearly utilizes domains and domain walls in its operation. Such a device makes use of the fact that films can be made in which it is clearly easier to move a domain wall than it is to nucleate or form the wall in the first place. Once a domain is formed it can be shifted around to various places on the film without generating new domains; in particular, it can be moved down a strip. The wiring pattern and pulse sequence shown in Figure 10 are that used by Broadbent. Each drive line winds back and forth across the sample. The "write" line just crosses the sample once; there is a similar "readout" line. Assume the film is magnetized vertically. A pulse in the "write" line of polarity shown creates a domain. This domain can be stepped along by alternately pulsing the drive lines and reversing the direction, as is indicated by the sequence.

---

9. Olson, C. D., and Pohm, A. V., Jour. Appl. Phys. 29 274 (1958)
10. Dietrich, W., Proebster, W. E., and Wolf, P., I.B.M. Jour. of Research 4 189 (1960)
11. Moore, D. W., Wescon Record 3 #4, P. 32 (1959)
    Moore, D. W., Proc. 1959 Electronic Components Conf., p. 11 (May 1959)
12. Broadbent, K. D., and McClung, F. J., Solid State Circuits Proc., p. 24 (1960); Broadbent, K. D., I.R.E. Trans. EC9 2, 21 (1960)

Of course it is not necessary to use a whole clock cycle to write if done at the appropriate time, as in Figure 10a.

The limitation of such a device is the ferromagnetic film. It must be possible to find a drive current which will extend a domain but will not create new domains; that is, it must be possible to make a film where the domain wall motion energy is clearly less than the domain nucleation energy. The separation must be large compared to the variation in the magnetic characteristics throughout the sample. When a better understanding of the films is obtained, the full potential of this device can be realized.

Ferrite devices reverse their flux in Region II by nonuniform rotation. The original device was a core in a matrix memory. For historical reasons Figure 11 should be included. It is a picture of a core array. The two drive lines and the readout lines are shown. This is just one of many different types of ferrite computer devices, such as aperture plates, sheets, transfluxers, coincident fluxers, inhibited fluxers, multiaperture (MAD) devices, biaxes, etc. Most all operate in Region II for their flux reversal.

Lastly there is Region III, where the reversal is very rapid. Cores operating in the impulse switching mode are probably reversing in Region III. By far the largest effort is in trying to develop a thin film memory working in Region III. Rubens and Pohm,[13] Bittman,[14] Raffel,[15] and Bradley[16] have reported various degrees of success. Except for the original try of Rubens and Pohm, most attempts have been to take advantage of the low demagnetizing factor in the transverse direction. Therefore, operation with a coincidence of two perpendicular fields has been attempted. Typically, the situation is as shown in Figure 12. The easy axis is as shown. Information is stored by magnetizing either to the right or left for a binary zero or one. Information is read by pulsing and observing a positive or negative signal. Bradley has introduced a clever scheme for simplifying the construction. He replaces one-half of the drive conductors with a conducting sheet. The eddy current image of the conductor completes the solenoid. For economic reasons it is desirable to make whole sheets of these spots at once. Such a desire puts a severe restriction on the nonuniformity that can be tolerated. Even so, it is now possible to buy film memory planes for small high speed applications.

It should be emphasized, in conclusion, that the new applications of magnetic materials mentioned here has been done for illustrative purposes rather than as a survey. No attempt has been made to include everything or even to obtain a balance in the kinds of things possible. An attempt has been made to indicate the class or type of device that generally operates in each of the three regions or by one of the three possible mechanisms of flux reversal to illustrate the mechanism.

---

13. Pohm, A. V., and Rubens, S. M., Proc. E.J.C.C. p. 120 (1959)
14. Bittmann, E. E., I.R.E. Trans. EC-8 92 (1959)
15. Raffel, J. I., Jour. Appl. Phys. 30 60S (1959)
16. Bradley, E. M., British Journal I.R.E. 20 765 (1961)

Figure 1

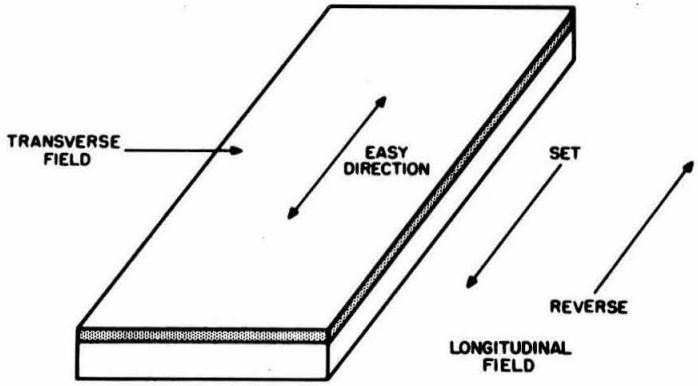
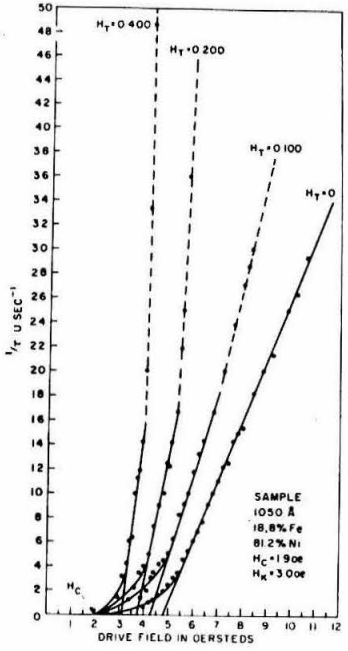


Figure 2



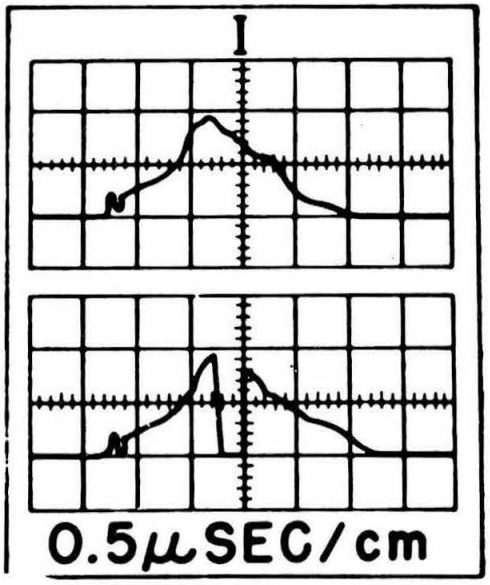


Figure 3

Figure 4

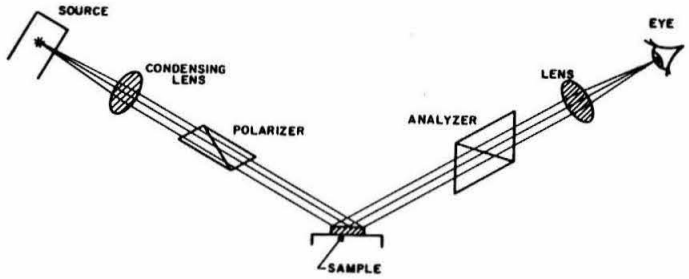
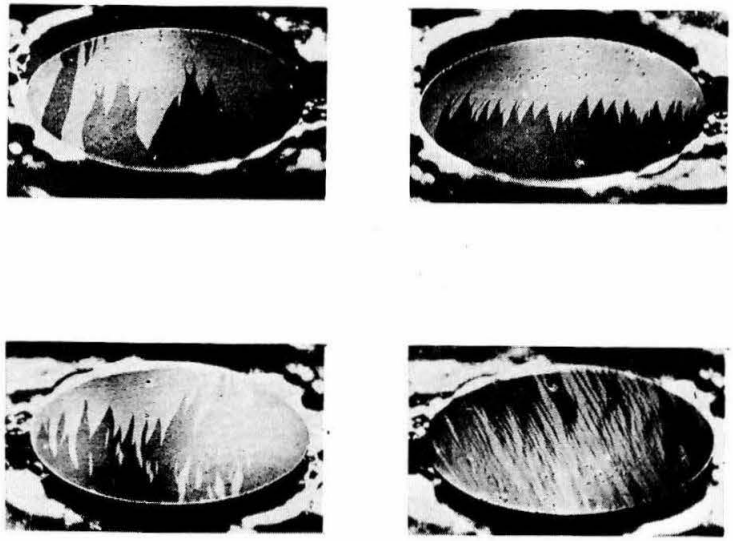


Figure 5



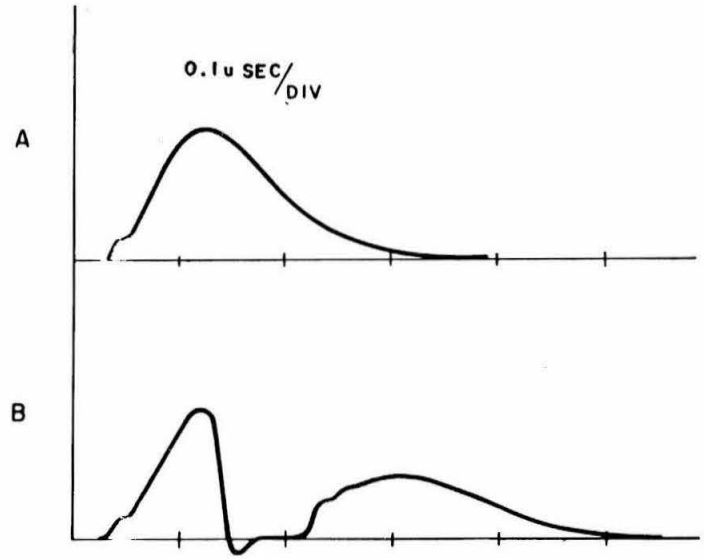


Figure 6

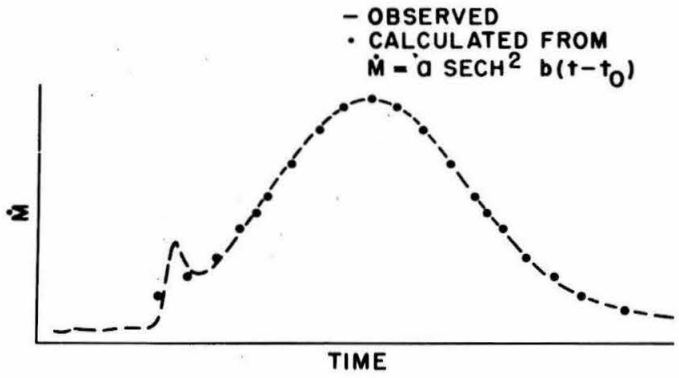


Figure 7

Figure 8

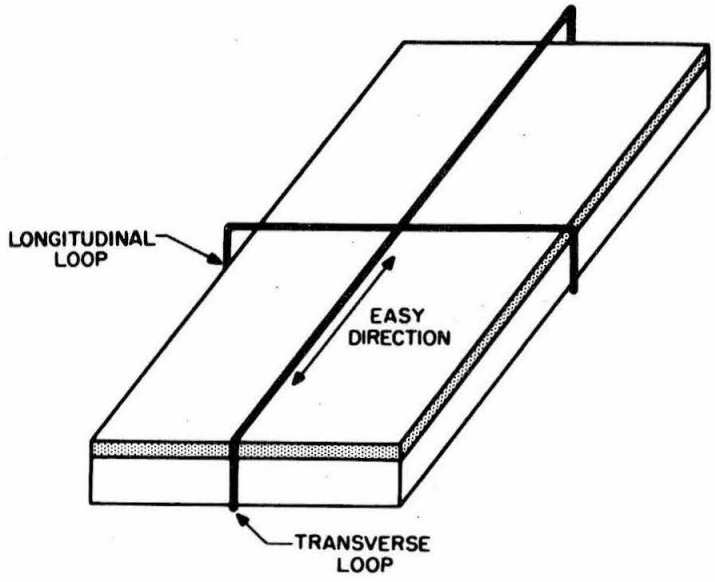
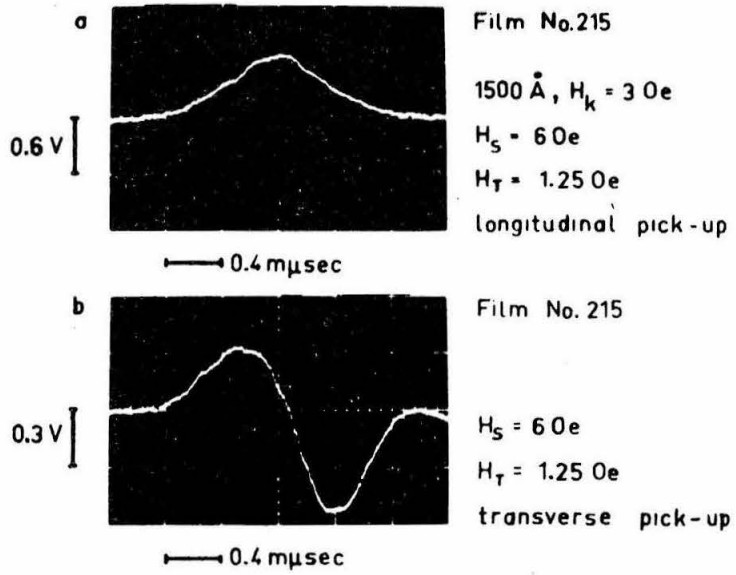


Figure 9



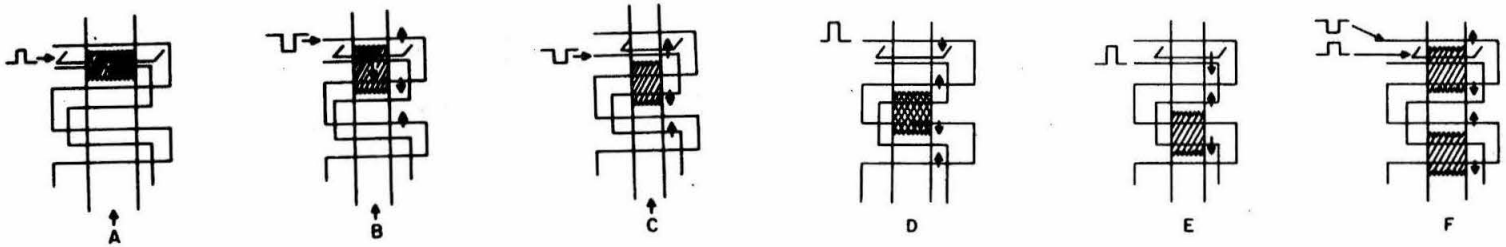


Figure 10

Figure 10-A

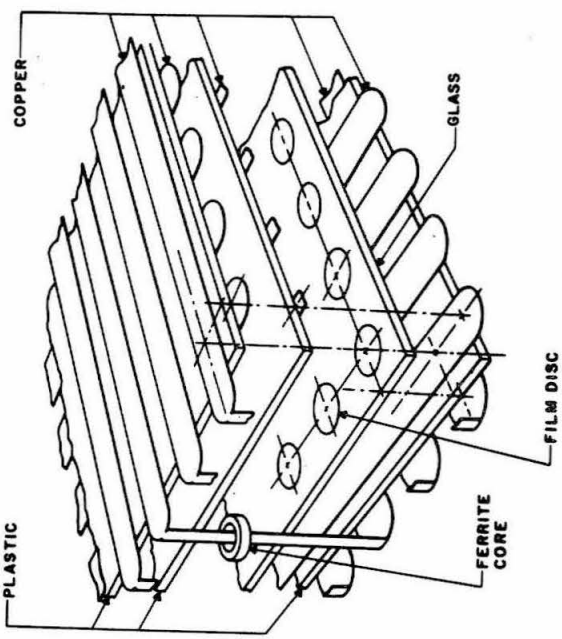


Figure 12

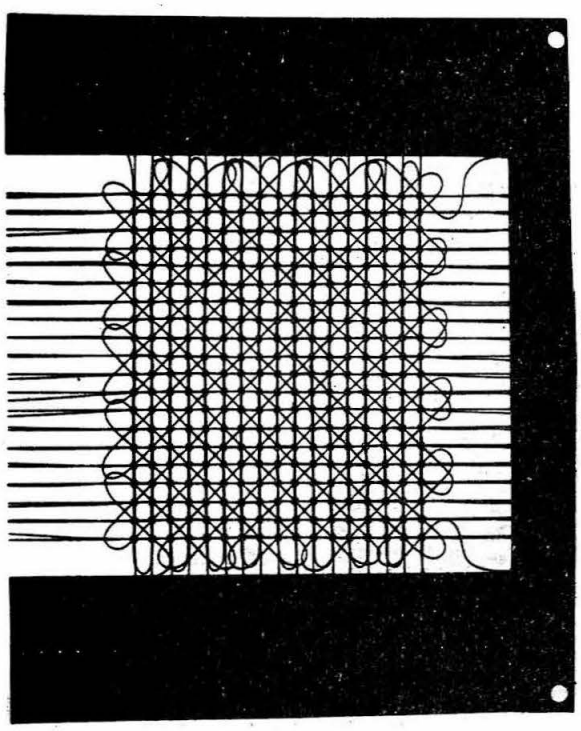


Figure 11

FERROELECTRIC DEVICES

J. Reid Anderson
Stanford Research Institute

Ferroelectrics exhibit certain remarkable dielectric properties which are in many ways analogous to the properties of ferromagnetic materials and this analogy is the sole basis for the term ferroelectric. In contrast to magnetism, ferroelectricity has been known for only forty years and the majority of useful materials have been discovered only in the past 15 to 20 years.

Most of the known ferroelectrics are ionic compounds which can be studied most easily in the single crystal form. Like semiconductors, many of the interesting devices have been fabricated from artificially grown single crystals. However, all of the materials in commercial use have been fabricated as polycrystalline ceramics of either single ferroelectric compounds or mixtures of compounds. Thin films of ferroelectrics have also been fabricated both by vacuum vapor deposition and by chemical deposition.

Figure 1 illustrates the phenomenal growth of the number of known ferroelectrics in recent years. Valasek is credited with making the first discovery of the ferroelectric phenomenom by observing a dielectric hysteresis loop in Rochelle salt in 1921. Important discoveries of other families such as KDP, $BaTiO_3$, GASH, TGS, the alums, and others occurred at later dates as shown. Most of the known ferroelectrics are structurally related to or are isomorphs of some of these families. The first antiferroelectric (not a magnetics man or someone who had tried to make ferroelectrics devices practical) was discovered in 1951 although they had previously been predicted by Kittel. Materials exhibiting the ferroelectric effect are certainly much more common than was originally suspected and if discoveries continue at the present rate more useful materials should appear in a few years. One word of caution however. While almost all of the over 80 ferroelectrics now known have contributed greatly to both our knowledge of ferroelectricity and to fundamental problems in solid state physics, they have not been good engineering materials for one or more of a variety of reasons such as solubility in water, or very low operating temperatures, or the structural limitations of small single crystals, or a lack of uniformity and stability of electrical properties.

Ten of the twenty crystal classes of piezoelectrics are pyroelectric. That is their internal electric dipole moment does not vanish and they are polar in nature. The dipole moment of the pyroelectrics change as a function of temperature. In addition to this effect in some pyroelectrics the dipole moments can be reversed more or less permanently. These crystals are the ferroelectrics. Thus all ferroelectrics are both reversible pyroelectrics and are piezoelectric but the converse is not true.

The spontaneous parallel alignment of electric dipoles in ferroelectrics or spontaneous polarization corresponds to the intrinsic magnetization in a ferromagnetic. This gives rise to the dielectric hysteresis loop shown in Figure 2 which shows spontaneous polarization, usually expressed in μ coulombs/cm^2, as a function of applied electric field. The loop shown is similar to those actually obtained with single domain single crystals of barium titanate and tri-glycine sulfate. The spontaneous polarization or hysteresis effect in ferroelectrics exists only below a certain temperature called the Curie temperature. Above the Curie temperature the crystals go into a higher state of symmetry which is paraelectric.

The structures of many of the ferroelectrics have been carefully studied but there is not yet a complete and well developed theory on ferroelectricity. Theories have been advanced by Devonshire, Slater, Matthias, Mason, and others. Qualitatively it is generally agreed that the spontaneous polarization in a ferroelectric is due to the anisotropic displacement of ions in the crystal lattice structure. For example, just below the Curie temperature, the Ti ion in $BaTiO_3$ is displaced from its central position relative to the oxygen octahedron surrounding it. The hydrogen ion moves between two stable oxygen bond configurations in KDP and the hydrogen nuclei also move in Rochelle salt.

Like magnetic materials, ferroelectrics have a domain structure, that is regions in which the electric dipoles are all aligned parallel. These domains are visible with polarized light in transparent crystals and may even be observed in motion through transparent electrodes when proper techniques are used. Powder pattern techniques employing colloidal suspensions in insulating organic liquids and acid etching techniques have also been widely used to observe and study domains in ferroelectrics.

In addition to dielectric hysteresis, a Curie temperature, and domain structures, ferroelectrics usually exhibit very high dielectric constants rising to a peak at the Curie temperature and falling off above that in accordance with the Curie-Weiss law. This is illustrated in Figure 3. It will be noted that several phase transitions in addition to the upper Curie temperature as shown by these discontinuities take place in $BaTiO_3$ as well as in other ferroelectrics.

Ferroelectrics also exhibit very large electro-optic Kerr effects and large dielectric nonlinearities. An example of the latter is shown in Figure 4 by the variation with D. C. biasing field of the small signal or reversible dielectric constant. Besides these useful phenomena, many other anamolous effects have been observed in ferroelectrics which have been the cause for considerable concern on the part of potential users. These include a tendency towards electrical and chemical instability in the polarized state, space charge layers near the surface of crystals, "fatigue" of polarization, bias of hysteresis loops, sensitivity of the hysteresis loops to ultraviolet radiation, the existence of strong Barkhausen pulses similar to those found in magnetics, and even electroluminescence.

With the wide range of compounds and crystal structures as are found in the various ferroelectric families it is not surprising that there would be a wide variation of physical characteristics as shown in Figure 5. Spontaneous polarization varies by a factor of nearly 500 to 1 between different materials. Coercive forces vary by at least 200 to 1 and Curie temperatures run all the way from -213°C to +700°C. As might be expected only a few of the ferroelectrics are useful at and above normal room temperatures, since most of the useful properties of ferroelectrics such as dielectric hysteresis and other dielectric nonlinearities occur either below or near to the Curie temperature. As can be seen the dielectric constants or permitivities can be quite high although most fall between 6 and 100. Like many of the other physical properties, the dielectric constants are strongly temperature and field dependent. Unfortunately some of the more stable ferroelectrics such as the mineral colemanite have Curie temperature so low that they cannot be considered for use with the usual types of electronic equipment. Some typical physical values for single crystals of $BaTiO_3$ and GASH are also given in Figure 6. The 60 cps hysteresis losses illustrate the very high internal losses in ferroelectrics when domain wall motions are involved as compared to some typical square loop magnetic materials. In experimental devices, such as a memory cell, the volume of ferroelectric materials may be only 1/1000 that of a 0.050 inch O. D. magnetic memory core so that the total energies required in each device may be quite comparable. However, in devices making use of the square hysteresis loop properties of ferroelectrics, it would be highly desirable to find materials with spontaneous polarizations as low as 0.1

micro-coulombs per square cm. and coercive forces in the range of 100 V/CM in order to reduce both internal heating and operating power requirements.

When voltage pulses having very short rise times are applied across the terminals of a ferroelectric capacitor, a current pulse due to the reversal of spontaneous polarization is observed as shown in Figure 6. The length of these switching current pulses, termed the switching time, varies as a function of applied field in somewhat the same fashion as you have just seen for magnetic materials. This is illustrated in Figure 7 for a typical sample of TGS. It will be observed that there is no threshold or cutoff field below which no reversal of polarization takes place as is found in magnetic materials. The absence of a fixed coercive force is also observable in ferroelectric hysteresis loops. As the rate at which a hysteresis loop is traced out is increased, the apparent coercive force increases.

At the very low applied fields, that is in this region, (lower left hand of Figure 7) there is evidence in $BaTiO_3$ crystals that considerable sidewise domain wall motion takes place when polarization is reversed. At somewhat higher fields the polarization reversal mechanism is one of domain wall movement either in a forward direction through the thickness of the crystal or sidewise where the wall moves in a direction parallel to the ferroelectric axis or a combination of both. At very high fields, switching times as short as 10 millimicroseconds have been observed in ferroelectrics indicating that at these high fields polarization may be reversed simultaneously rather than sequentially by domain wall movement as happens at low fields.

Ferroelectrics have been in widespread commercial use in polycrystalline ceramic form both as miniature capacitors and as electromechanical transducers. Mixtures of barium titanate and strontium titanate are used for capacitors and ceramic forms of barium titanate and of lead zirconium titanate are widely used in electromechanical transducer applications.

A number of other interesting device applications of ferroelectrics are listed on Figure 8. These are all applications which have been tried at least on a laboratory scale.

The first group of applications at the top of the figure, excluding the blocking capacitor, all exploit dielectric nonlinearity but do not make direct use of reversing spontaneous polarization or of dielectric hysteresis. In fact, dielectric hysteresis loss is of a distinct disadvantage.

Dielectric amplifiers based on using the non-linear properties of ferroelectrics to modulate carriers or the use of controllable impedances similar to those used in magnetic amplifiers have been studied since 1948. There are reports that such amplifiers came into widespread use in Russian home radio receivers and the Russian technical literature is filled with references to new nonlinear ferroelectric ceramics termed Varikond I and II types developed for such applications. They claim both voltage and power amplifiers in multistage form which work over a wide frequency range. The dielectric amplifiers constructed in this country have a possible upper frequency limit of 10 mc and were never competitive with semiconductor amplifiers because of a lack of versatility and requirements for high frequency power supplies. Japanese have carried on further work in this field with parametron circuits for their computers but these have not been competitive with magnetic parametron circuits.

The next group of applications (in Figure 8) depends on spontaneous polarization reversal effects or more on the memory and charge quantization ability of ferroelectrics.

The potentially most important applications here are for memories and shift registers. Shift registers up to 20 stages in length have been constructed with ferroelectrics and registers have also been constructed with combinations of ferroelectric and magnetic devices. The chief virtue of such registers is that they are relatively easy to design and construct and require low currents at reasonable operating speeds, i.e., 5 to 10 kc. Again, however, they suffer from lack of cheap, readily controlled materials and must use avalanche diodes to provide threshold fields.

The first energy conversion application is based upon the temperature dependence of dielectric permitivity near the Curie temperature. Energy converters of this type have been studied by Sigmund Hoh, of I. T. & T. They are handicapped with presently available materials by conversion efficiencies of less than 1%. The chemical to electric energy converters depend partly on the piezoelectric properties of the ferroelectrics as they make use of shock waves generated by explosive detonations to generate high voltage pulses. This is a very promising - although limited - use of ferroelectrics and some critics might say finally we've found something worthwhile for ferroelectrics - blow them apart!

Digital memory and the use of ferroelectrics with electroluminescent displays will be briefly described. At present the microwave devices are limited only by low Q's and temperature sensitivity but material improvements might remedy this. The microwave blocking capacitors have extremely wide bandwidths (2 mc - 10 Kmc) and appear to be commercially attractive.

The possibility of using the rectangular hysteresis loop characteristic of ferroelectrics for digital computer storage and certain logic circuits was probably partly responsible for stimulating much of the search for new ferroelectrics after 1950. A typical scheme for a coincident voltage memory device is shown in Figure 9. The planar geometry of ferroelectrics offers the possibility of placing many elements (here 256) on one single slab of material and of storing by applying coincident voltages. Small ferroelectric memories of this type have been operated in the laboratory both with conventional pulse sensing and with non-destructive sensing of stored data at each of the crosspoints by detecting the phase of difference frequencies generated when two different interrogating frequencies are applied to the matrix. While they are potentially more economical of storage space and potentially cheaper than magnetic memories, commercial versions have not been achieved because of instability of the available ferroelectric materials and the lack of threshold fields which complicates the accessing problems and limits the size. Were these problems solved, one would still have to examine the access and reading circuitry necessary to ascertain if these would become competitive with the very successful magnetic memories. A typical memory storage of this type on a single crystal of $BaTiO_3$ is shown in Figure 10. With more stable and less expensive materials it is highly probable that ferroelectrics could offer a very compact and low cost memory where small amounts of data, i.e., 10 to 1000 bits have to be stored for a short time. Thin films of ferroelectrics also offer a very promising approach to small memories. It is highly improbable that ferroelectric memories will ever compete speedwise with magnetic memories (unless extremely thin films can be fabricated) because the switching mechanism of the former depends upon the physical movement displacement of ions as opposed to the reversal of electron spins in magnetics.

The next application I would like to discuss is one which is in pilot plant production and which illustrates a use of ferroelectrics to control an electroluminescent display. It is also interesting that it makes combined use of three different solid state devices - electroluminescent display panels, ferroelectrics, and silicon diodes. I am indebted to Dr. E. A. Sack, of Westinghouse, the originator of this application for supplying these illustrations. The objective is one of constructing a large flat display panel 4" by 8" having 256 separate elements per square inch. Each of these elements must be

separately selected and controlled in brightness in order to provide a means of imaging video, digital data, and other forms of pattern information. An over-all schematic of the construction of this system is shown in Figure 11. The letters ELF spelled out by the activated electroluminescent rectangles stand for Electroluminescent Ferroelectric Display.

The excitation voltage applied across each segment EL of the electroluminescent panel, as shown in Figure 12, is determined by the capacitance of the associated portion of the ferroelectric control structure. This capacitance is, in turn, altered by the control bias of a charge pattern distributed over the screen and applied for this segment between the junctions of B and A. Since the ferroelectric is a good insulator, the control charge pattern, and hence the image on the electroluminescent panel, remains constant until it is purposely modified. Changes in the voltage across EL of the order of 3 to 1 due to capacitance changes will give changes in brightness ratios of the order of 50 to 1. More recent development in nonlinear dielectric materials has produced some whose unsaturated to saturated capacitance ratios are as high as 20:1.

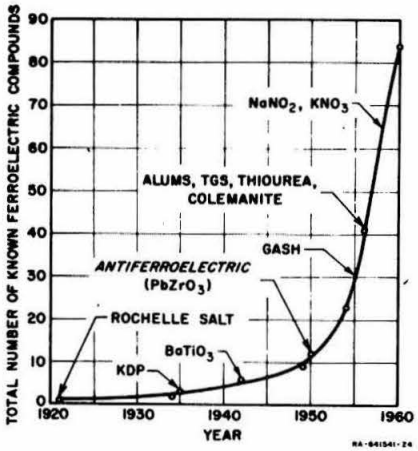
It is apparent that ferroelectrics have been going through the initial search phase both as materials and as devices. All kinds of experimental materials have been synthesized and devices and circuits have been devised in the laboratory for almost every type of electronic application.

We should expect to see more in the development of better new materials in the future rather than a continual flood of exotic new ferroelectric compounds. It is hoped that materials will be developed which are close to the silicates in stability and techniques advanced for making chemically deposited thin films as well as epitaxially grown ferroelectrics. Almost nothing has been done with small single ferroelectric particles imbedded in insulators or conductors - another fruitful path for future research. Perhaps we could also find more new ferroelectric minerals in nature which would have useful properties. The materials which are both ferroelectric and ferromagnetic are almost completely unexplored.

As storage and logic devices, ferroelectrics have suffered as direct competitors with magnetic devices because of three basic limitations - the lack of a fixed threshold field, the two terminal nature of the device, and the fact that they cannot provide impedance transformation or anything equivalent to a turns ratio. Recently some anti-ferroelectric crystals in the sodium niobate class have been found by Pulvari to exhibit something akin to the threshold effect found in magnetics. Materials with similar properties have also been reported in England and Russia. Although this is a very promising development, these types of materials are not well understood and are still in the early experimental stage. The favorable geometries of ferroelectrics for multielement units, their compatability with microminiaturization and their relatively large signal outputs would make them attractive for a few special digital circuit applications if we could obtain better materials.

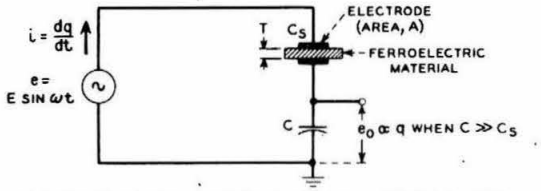
There seems to be no question that the nonlinear dielectric properties of ceramic ferroelectrics will continue to make them useful in many device applications particularly in the microwave field and for control of other devices whose impedance gives a good match such as electroluminescent panels. A number of new device applications which have not yet been explored may also become important. These are electro-optic transducers, new recording techniques, combinations of magnetic and ferroelectric devices, and the newer type of computer device concepts such as the neuristor and the perceptron. Ferroelectrics have had an interesting and sometimes disappointing past when they were applied prematurely to the wrong circuits but they should have a much more useful future as they come of age.

Figure 1

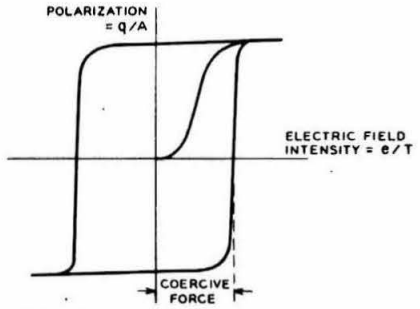


RATE OF DISCOVERY OF FERROELECTRICITY IN COMPOUNDS

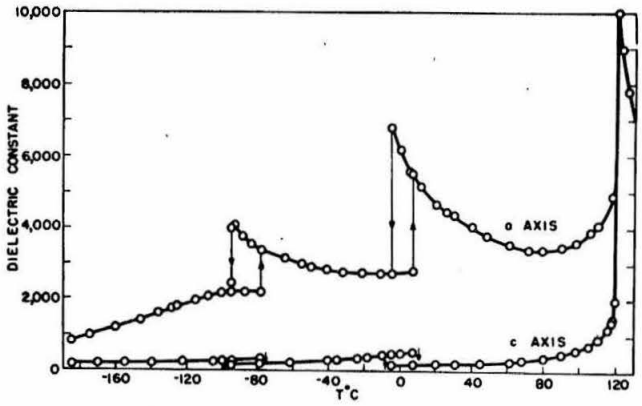
Figure 2



METHOD OF OBSERVING FERROELECTRIC HYSTERESIS LOOP

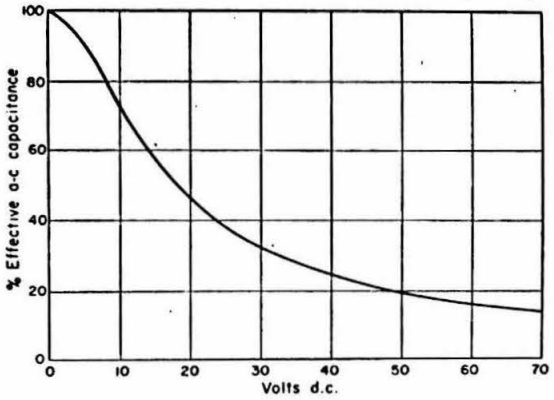


FERROELECTRIC HYSTERESIS LOOP



TEMPERATURE DEPENDENCE OF THE DIELECTRIC CONSTANTS OF $BaTiO_3$ (After Merz)

Figure 3



CHANGE OF A-C CAPACITANCE VS D-C POLARIZING VOLTAGE FOR $BaTiO_3$-$SrTiO_3$ CERAMIC; K' AT ROOM TEMPERATURE = 8000

Figure 4

Figure 5

RANGE OF FERROELECTRIC PROPERTIES

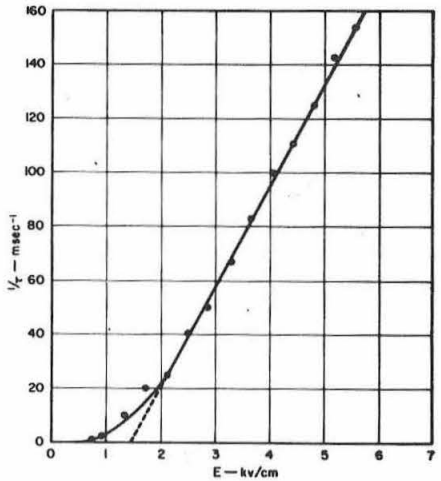
	MINIMUM	MAXIMUM
SPONTANEOUS POLARIZATION	0.06 $\mu C/CM^2$	30 $\mu C/CM^2$
COERCIVE FORCE AT 60 CPS	0.15 KV/CM	30 KV/CM
CURIE TEMPERATURE	-213°C	+700°C
REVERSIBLE DIELECTRIC CONSTANT ϵ_R	3	5000

TYPICAL VALUES

MATERIAL	SPONTANEOUS POLARIZATION $\mu C/CM^2$	COERCIVE FORCE KV/CM	HYSTERESIS LOSS 60 CPS ERGS/CM^3/CYCLE	ϵ_r
$BaTiO_3$	26	0.75	2.7×10^5	≃ 60-5000
GASH	0.33	2.2	1.5×10^4	≃6

(MAGNETIC MATERIALS)

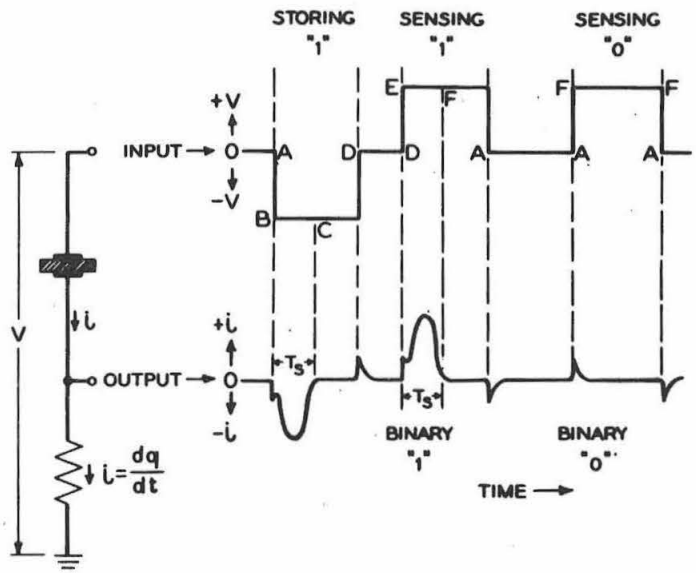
TYPE S-1A Square Loop Ferrite	0.76×10^3
4-79 Molypermalloy	0.13×10^3



RECIPROCAL OF THE SWITCHING TIME τ AS A FUNCTION OF FIELD FOR TGS

Figure 7

Figure 6



APPLICATIONS OF FERROELECTRIC DEVICES

CONTROL FOR ELECTROLUMINESCENT DISPLAYS

DIELECTRIC AMPLIFIERS
VOLTAGE TUNABLE OSCILLATOR ELEMENTS
PARAMETRONS AND FERRORESONANT CIRCUITS FOR COMPUTERS
VOLTAGE TUNABLE FILTERS
D. C. BLOCKING CAPACITORS FOR MICROWAVE CIRCUITS
MICROWAVE PHASE SHIFTERS, HARMONIC GENERATORS, IMPEDANCE MATCHING NETWORKS, ATTENUATORS, AND PARAMETRIC AMPLIFIERS

INFORMATION STORAGE
LOGIC CIRCUITS AND SHIFT REGISTERS
COUNTING CIRCUITS
PULSE TRAIN GENERATORS
FREQUENCY MODULATION DETECTORS
A. C. CURRENT REGULATORS
TRANSCHARGERS OR TRANSPOLARIZERS

THERMAL ENERGY CONVERTORS
CHEMICAL TO ELECTRICAL ENERGY CONVERTORS

Figure 8

Figure 9

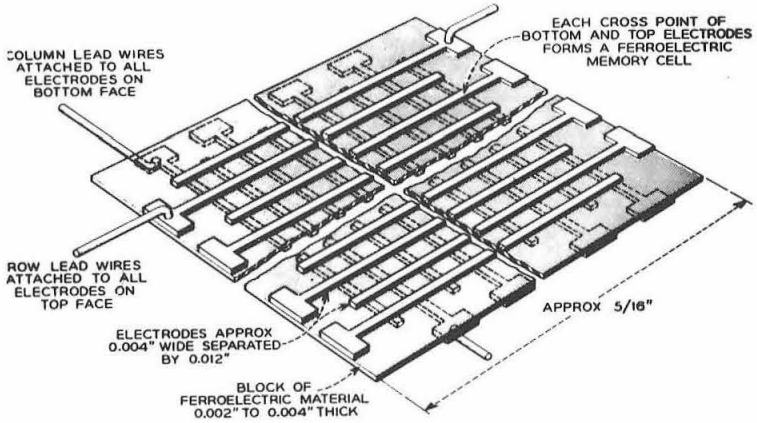
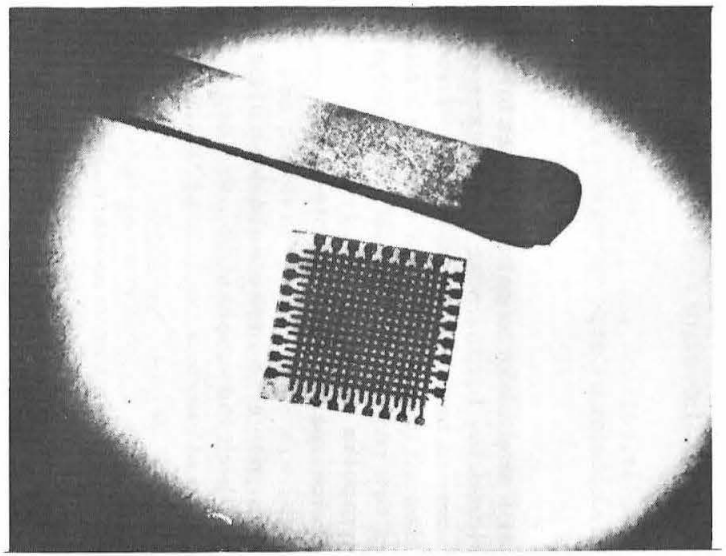


Figure 10



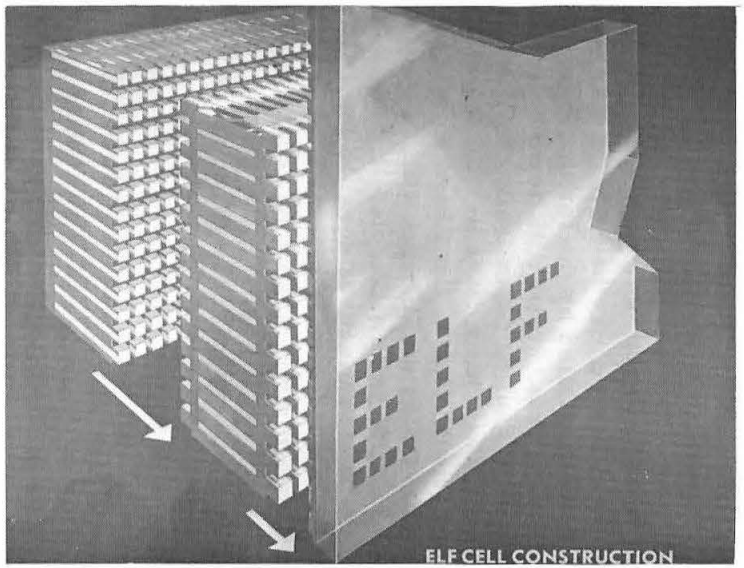


Figure 11

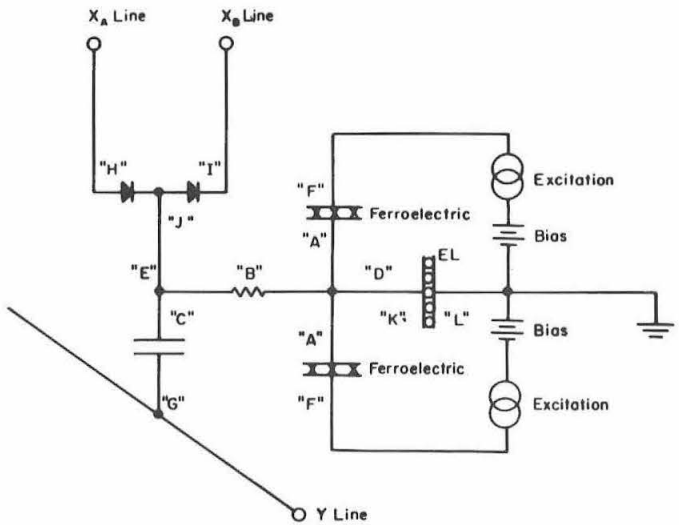


Figure 12

THERMOELECTRIC DEVICES

F. E. Jaumot, Jr.
Delco Radio Division, General Motors Corporation

If I may jump into the middle of my talk first I would like to point out that for those of us who have followed thermoelectricity work closely for the past eight years or so, the most dramatically visible change has occurred in the last two years. Two years ago if one attended a conference on thermoelectricity he found the vast majority of the papers to be on materials with perhaps a few papers on some of the difficulties associated with device design. Recently, symposia on thermoelectricity have emphasized actual hardware, device design, device operation, and potential.

Actually, I do not believe, and I do not think anyone else in the field does either, that the major problems have changed. Rather, I think the emphasis on devices is the natural result of improvements which have been made on a continuing basis with little indication of diminution.

Speaking of improvements, obviously we could use another breakthrough of the type provided by the emergence of semiconductor technology. At the same time, I am firmly convinced that the eventual commercial success of the field does not depend on achieving another such dramatic breakthrough.

With that rather inverted introduction, I would now like to describe very briefly what I am talking about.

Our time will not permit much of a discussion of the basic thermoelectric effects -- and, incidentally, any phenomenon involving an interchange of heat and electrical potential energy may be called a thermoelectric effect. Figure 1 lists the effects commonly implied by the term. In particular, I want you to note the expressions giving the rate of heat transfer due to the Peltier and Thomson effects. The coefficients π and N depend only on the temperature and on the material of a junction of dissimilar conductors in the case of the Peltier effect or on the material of a single conductor for the Thomson effect. S, the Seebeck coefficient, is the name given to the rate of change with temperature of the sum of emf's due to thermoelectric effects in a complete circuit. Then, since S is intimately related to π and N and, in fact, is equal to the Peltier coefficient divided by the temperature, we see that the entire potential for heating, cooling, or generation of electricity by thermoelectric means which is indicated by these expressions is determined by the materials used, assuming temperature is at our disposal. Inasmuch as we can express any one of these parameters in terms of the others, we will be concerned only with the Seebeck coefficient.

Since the potential of thermoelectricity rests on materials parameters, one wants to know which ones are important and the relations among them. To determine this, the obvious thing to do is to write down the expressions for the efficiencies or the coefficients of performance of the devices one wants to use. Doing this one gets the results shown in Figure 2 for the three principal applications. Here T_1 is the temperature of the hot junction, T_0 the cold junction temperature, S the Seebeck coefficient, σ the electrical conductivity, and K the thermal conductivity.

There are several things we want to pick up from this figure, but first I should point out that they are approximate expressions, derived neglecting the Thomson effect

and optimized with respect to geometrical factors and the ratio of load to internal resistance. For present materials, they are remarkably good.

The important thing for our discussion is the fact that in each case we want the factor M as large as possible. But, in terms of the materials, this simply means that we want z, which we call the figure of merit, to be a maximum. We will concentrate on this figure of merit from here on since it is the only factor in which the materials parameters appear and consequently it completely provides our materials description. Apart from questions of melting, volatility and mechanical properties, it also completely determines what we can do in application. Further, since we lose nothing of the physical principles by so doing, we will concern ourselves only with the figure of merit for a single material.

Obviously, we would like to have maximum values of S and σ and a minimum value of κ. The problem is that these are not entirely compatible as we can see from Figure 3. The expressions on this figure are the simplest expressions we can get for non-degenerate semiconductor theory and are far from satisfactory for serious consideration. However, they do illustrate the rough principles which hold. Here e is the electron charge, μ is the mobility of the charge carrier, n is the number or density of charge carriers, k is Planck's constant, T the temperature, m_s the effective mass of the charge carrier. Also of great importance, as shown both by the equations and Figure 4, is the fact that the thermal conductivity can be expressed as a sum of the phonon or lattice thermal conductivity and the thermal conductivity due to the charge carriers.

As can be seen from the plot of the various factors as a function of the number of free charge carriers, the electrical conductivity is roughly proportional to n. On the other hand, S tends to zero as n tends to infinity and S approaches infinity when n approaches zero. The lattice thermal conductivity is independent of the number of charge carriers and the electron component is proportional to n to a first approximation.

Thus, there is an optimum value of the carrier concentration. Using the simplest form of the theory, the product $S^2\sigma$ is a maximum at concentrations, n, of the order of 10^{18} to 10^{20} charge carriers per cubic centimeter. This is approximately 1000 times smaller than for metals and 1000 times greater than for the more common semiconductors forcing one to conclude that appropriately doped semiconductors are the most likely candidates for materials.

These expressions tell us two things about the charge carriers. First, they should have maximum mobility and second, their effective mass or, if you will, density of states mass should be as large as possible. In a very rough way, a large effective mass means a complex band structure.

The one thing that is completely clear from these expressions is that we want the phonon contribution to the thermal conductivity to be a minimum.

Since the validity of the figure of merit as an indication of the usefulness of a material in a practical application is well established, we see that we have fairly good empirical guides to the desirable materials. The problem, of course, is that because these guides pretty much eliminate elemental semiconductors, one is faced with so many possibilities, with nearly every possibility involving three or more elements. This leads to so great a number of combinations and variations of combinations that the achievement of major success by cut and try is certain to be very laborious and extremely expensive. Thus, there is badly needed better fundamental understanding for predictability. We shall try to enlarge on this when we discuss the future. For

the moment, however, I would like to look at the present status and the progress that has been made to date.

Figure 5 shows the improvement since Peltier made his remarkable discovery. The big breakthrough, of course, came with the advent of semiconductor technology and the metallurgical techniques making possible the fabrication of "tailored" materials.

The best values of z we have today run between three and four times 10^{-3} with units of reciprocal degrees. But what do these values mean in terms of something we could sell? This is answered in the next two figures for the two most popular applications, generators and refrigerators. For example, in Figure 6, a generator with a z of 4×10^{-3}, a hot junction at T= 600°K, and a cold junction at room temperature would have an optimum efficiency of 15.5% which would put us in business in many applications. Unfortunately, we need both p and n type materials with this high value of z. We also need to know how to construct devices so that we do not lose a significant portion of the materials figure of merit. To date, small generators with over-all efficiencies of 10 per cent have been fabricated relatively reproducibly, but most reasonable useful ones, in terms of size, have been between 5 and 10 per cent efficient.

For refrigerators, shown in Figure 7, we see that our z of 4×10^{-3} would mean a coefficient of performance of about 1.3 with a Δ T of 30°C and a hot junction temperature of 300°K. If we could use a hot junction temperature of 500°K, we could operate a refrigerator with a Δ T of 30°C with a coefficient of performance of 4.5 which is good indeed. Incidentally, this points up the importance of knowing the hot junction temperature if one is talking about operational efficiencies.

It is a little difficult to compare compressor type and thermoelectric refrigerators directly, but in terms of usage, a thermoelectric refrigerator, without freezer, operating over a temperature difference of 30° C from 300°K is roughly comparable. A comparable commercial household refrigerator, without freezer, will have a coefficient of performance of about 1.1. However, it is not the whole story in terms of speciality devices, or even some not so special when size is concerned. This is shown in Figure 8, although not to the best advantage. Incidentally, the thermoelectric calculations here were made in 1958 and are based on a figure of merit of 1×10^{-3}. Of course, we can do better today and if we could build a device with $z = 3$ or 4×10^{-3}, we would move this curve down significantly. The important thing here however is that thermoelectric device costs are almost directly proportional to their size until we get into very large devices where complex cooling problems arise. On the other hand, vapor compression machines do not lower in price very much as one goes smaller and, in fact, I am told they actually increase in price for very small capacity items.

This brings us to the specific question of just what can we do today. My favorite answer is, almost anything that can be afforded; and even with cost included, there are several areas in which thermoelectrical devices are practical today, in the strictest sense of the word. These are shown in Figure 9. The common characteristics of these applications are a need warranting a premium cost or a need that cannot be met adequately with standard methods, coupled with light loading and moderate temperature ranges.

I am sure you are all aware of the fact that a great many speciality devices falling into these areas have been built for feasibility and display purposes; in addition there are a fair number on the market, notably by Westinghouse. Recently they added to their thermoelectric product line four generators for industrial applications with ratings of 5, 10, 50, and 100 watts and costing from $1700 to $6500. Incidentally, efficiency ranges from 2 to 9 per cent depending on use.

Many serious studies are underway to push the applications frontier back appreciably. All in all, however, the larger scale applications are being studied at present primarily for space and navy applications.

In the space field, I am familiar with a study comparing photovoltaic cells, thermionic converters and thermoelectric systems. A thermoelectric system with thermal heat storage turned out to be lightest by at least a factor of two for equivalent power output. As a matter of fact, it will be possible to construct solar converters weighing as little as 40 pounds per kilowatt.

In the final analysis, the Navy has supplied the real impetus to thermoelectric applications. This is true for both generation and refrigeration. In the isotope powered class of generators we have the Snap series ranging up to 250 watts. In the fossil-fuel fired class, Westinghouse has delivered to the Bureau of Ships a water cooled, 5 KW generator which operates at 4.7 per cent efficiency. Since the design was frozen some time ago, Westinghouse now feels they could do better than 9 per cent and cut the size and weight in half too.

Summarizing the generation application, the latest NRL status report indicates that prototypes and feasibility studies, well underway, cover the range in power from 5 watts to 10 megawatts.

As important as generation is to the Navy, perhaps someday even for a ship propulsion, refrigeration and, particularly, air conditioning for submarines are probably more important. Since it appears that a thermoelectric air conditioner system would require only 30 % of the space required for conventional compressor systems, fewer spare parts will be required and operation will be much quieter, it is not surprising that there is considerable activity in this area.

Several companies are working on prototype air conditioners up to one ton in size, but perhaps the most interesting device is a three-purpose unit Westinghouse is constructing for the Navy. It includes a one-ton air conditioning unit for supplying chilled water to a standard Navy space cooler; a central thermoelectric heater for supplying hot water to a standard Navy space heater; and a 2 cu. ft. refrigerator-freezer capable of maintaining zero F continuously.

All in all, these items and many others would appear to indicate a rosy future for thermoelectricity. There are problems, however. All the materials problems are far from solved and the imminence of feasible systems creates an urgent need for work on these. "Unfortunately," and I am now quoting from the previously mentioned NRL Report, "there is no immediate prospect for initiating these studies. A request for funds for this work, which was initially encouraged and technically endorsed, was ultimately refused for policy reasons. It does not appear wise to move ahead rapidly on devices without the necessary supporting research on materials. The availability of quiet maintenance free power supplies for space, for ship propulsion, ASW or any other application, has therefore been set back a substantial period of time."

I think I have delayed long enough getting into materials. However, I believe Figure 10 summarizes most of the systems of greatest interest at present with the hasty disclaimer that no attempt has been made to include all those which are promising. In the first column we list the general systems and indicate the most studied combinations. In the second column we have included the temperature to which these systems can be used because the field is getting a little desperate for high temperature materials for power generation. However, in this connection I might point out that although, for power generation we would like almost as high a temperature as we can get, the attractive posibilities of focused solar and nuclear heat sources fall in the range from

500 to 1000°K. Below about 600°K we have to consider the material primarily useful for refrigeration.

The third column gives the best results obtained in a given system, which appear to be reasonably confirmed. That is to say, there have been a number of startling announcements in the last five years which have not been confirmed. A recent case in point was the announcement by the Nuclear Corporation of America at the American Rocket Society Space Symposium that they had prepared gadolinium selenide (Gd_3Se_4) with a figure of merit of 45×10^{-3}. So far, no one appears to have confirmed or reproduced these results.

Virtually any of these materials with high z values have doping beyond that indicated here. For example, a base Bi_2Te_3 material may be doped with Bi_2Se_3 and CuBr or iodine for more effective n-type conductance.

One point that is a little disheartening, Bi_2Te_3 and PbTe, and particularly Bi_2Te_3 are still the best materials we have. Figures of merit of nearly 2 were achieved in Bi_2Te_3 almost six years ago. However, as I mentioned earlier, in terms of a widening scope to the field, encouraging progress is still being made.

Finally, a striking fact is that all of the good alloys and compounds involve Group VI materials and particularly tellurium which is not readily available at low cost. To be sure, the rare earth sulfides so far appear to be best but more work could easily shift this to tellurium and the general desirability of heavy atomic weight elements would indicate it probably will. As for the three-five compounds with their very attractive high mobility, I am inclined to discount these because I am not convinced that the critical factor of thermal conductivity can be licked. In fact, at present they exhibit reasonable z values only near the upper end of their useful temperature range where the thermal conductivity appears to decrease.

We mentioned earlier that we had reasonably good guides to materials -- and the materials in Figure 10 confirms this, at least in part. As a background to our discussion of the future direction of thermoelectricity and the immediate problem of a maximum value of the figure of merit, I would like to review the initial requirements in slightly different words.

From our discussion of z we said we want as high a Seebeck coefficient as is compatible with optimum charge carrier density and we want a maximum ratio of electrical to thermal conductivity; this means we need relatively few charge carriers having very high mobilities and a minimum lattice thermal conductivity.

Physically, this translates to strong scattering in the phonon system to reduce lattice thermal conductivity and weak scattering in the electron system to obtain large mobility. Also, the material should be extrinsic and have low crystal symmetry.

Once one has a good starting material, there are several things he can do to improve it. First of all, one dopes the material to obtain optimum carrier concentrations. Once this is done, the ratio of σ/κ can usually be further improved by reduction of the lattice thermal conductivity. This is probably best achieved by introducing into the lattice another substance, either element or compound, which crystallizes in a similar lattice and has approximately the same lattice constant. The distortion of the basic lattice by the added impurity is then relatively small and is limited to crystal regions in direct contact with impurity atoms. Such distortions are reasonably effective in scattering the short wavelength thermal oscillations, but since the lattice periodicity is not greatly affected, the electron waves with their longer wavelengths are not effectively scattered and the current carrier mobility is not affected significantly.

On the basis of present information, how good can we expect materials to be? Is there a maximum figure of merit? There has been a lot said about this subject and there are a lot of ways to look at it. Actually, it is possible to construct hypothetical models which would exhibit arbitrarily large values of the figure of merit. However, if we make the assumptions necessary to the calculations compatible with the interdependence of S, σ and κ one finds, more or less, for present materials, one can calculate approximate bounds to the figure of merit. The values one gets range from 2×10^{-3} deg^{-1} to 20×10^{-3} deg^{-1} depending on whether the lattice thermal conductivity is much greater, approximately equal to, or much less than the electron thermal conductivity. It would appear that to exceed 6 or 7 times 10^{-3} one may need a new mechanism to reduce lattice thermal conductivity without decreasing carrier mobility. That is, in terms of present materials, to achieve an order of magnitude improvement in z we need either an order of magnitude reduction in lattice thermal conductivity or an order of magnitude increase in carrier mobility or some combination of both.

I have spent this time on rather unconvincing arguments concerning the maximum figure of merit primarily to show that the major advances in this field will come largely through a better understanding of transport phenomena in our more contrary solids -- insulators, semiconductors and semi metals.

Transport phenomena, particularly in the higher temperature ranges, are remarkably little understood at present; in fact, I do not have to tell you that the uncritical use of simple semiconductor theory, with which we are all so familiar, can lead to large errors in the estimation of even the simplest parameters when the physical situation becomes the least bit complex.

This brings us to the future and what do we need to do to realize it? I am as convinced today as I was eight years ago that there is a great future in thermoelectricity -- both technical and economic. What we need is more basic knowledge. We need theoretical work very badly, coupled with supporting fundamental research on materials; these must proceed together on a logical basis. At the same time, we are in a position to continue materials development and studies of device design criteria on the basis of our present knowledge. Let me cite a few examples in more or less reverse order.

The literature is full of design criteria for devices and aids to designing devices -- computing programs have even been developed -- this is all to the good. On the other hand, the problem of low resistance contacts dominates the size and weight of a device but so far, much of this work has been on the roughest kind of cut and try basis. The diffusion of the materials used to make contacts, as well as the diffusion of the doping agents used, determines the operating life of a couple. We need basic information and we need the logical extension of it, namely, how to arrest or control diffusion in a favorable way. So far, our best materials are very brittle; in fact, they have unfortunate mechanical properties all around. We need studies in all areas of mechanical properties. Finally, we have had very little work on performance of our materials except under the most benign environmental conditions. We need to know how the devices work under use conditions and the limits to which we must go in packaging them so they are useful. Once we know more about these factors we can study fabrication techniques more intelligently and look for methods for quantity production of the materials as well as the necessary cost reduction.

In the materials development area we need, first of all, more detailed studies on systems which appear to have desirable thermoelectric properties other than the IV-VI and V-VI systems. Since nearly all the best materials to date have been ternary or quarternary systems doped with still other materials, we need a method of predicting optimum composition, doping level and the probable maximum figure of merit from measurements on one or a few samples. The process of finding these conditions by cut and try is much too laborious and expensive. I am told privately that a computer

program has been written which appears promising; I sincerely hope so. As I mentioned, it has been shown that we can reduce the thermal conductivity by alloying; now we need confirmation of the evidence of Lawson in this country and Elpatevskia in Russia that mobility may also be increased by alloying. Particularly we need to know whether this is in any way a general possibility. More prosaically, we need positive evidence of the importance of extreme chemical homogeneity and perfect physical texture. And there are many other materials development questions.

It is in the theoretical and associated fundamental work that we start farthest back and perhaps have our greatest need.

What we really need is to set up the general transport problem with all the electrical, magnetic, and thermal effects and solve it. The trouble is that it is almost impossible to do. I will not try to impress you with the progress that has been made nor harangue you with what I think is wrong with what has been done; I have discussed that elsewhere. Suffice to say, much remains to be done and I would like to point out just a few examples of where understanding appears to be vital to the future of thermoelectricity.

First of all, the thermal properties are in bad shape. In spite of the excellent work of Krumhansel and others, we have no exact theory of thermal conductivity. One specific item -- the one thing almost everybody agreed on as to thermal properties was that a high melting point inevitably meant a high thermal conductivity. Yet Westinghouse had reported a thermal conductivity as low as 0.005 watt/cm deg C for samarium sulfide with a melting point above 2000°C. If this and similar data can be confirmed, it badly needs explaining since if we have to live with the concept of a high thermal conductivity in high melting point materials, we face a severe restriction in efficiency of high temperature thermoelectric devices.

Another area that needs work is the energy gap. In elemental semiconductors, the values for the energy gap obtained by optical and thermal means agree quite well, but they do not in the heavy compounds. Excess charge carriers will not explain the difference and although defect structures might, we need to know more of the details, particularly as to what the meaning of this departure is in terms of other properties.

In a more general area, we might mention that the familiar parameters including energy gap, effective mass, dielectric constants and mobility, although exhaustively studied and discussed, still are without any simple theory that can describe their change with temperature and other external forces. There are a number of empirical rules, but why are the interesting lead compounds an exception to nearly all of these rules? Actually, the positive energy gap dependence on temperature, the $T^{5/2}$ dependence on mobility, and the virtually constant energy gap in the homologous series of lead compounds are very desirable for thermoelectricity, making it even more important that we understand them.

I could go on, but I believe I have cited enough examples of the understanding we need to supply substance to my favorite closing statement. Whether we ever see large scale usage of thermoelectric refrigerators or generators or not, and I believe we will, this field will certainly give us materials which will find applications in many fields of technology. But perhaps the most important effect the interest in thermoelectricity will have is the increased understanding of solid state physics. And there lies a future in itself.

Figure 1

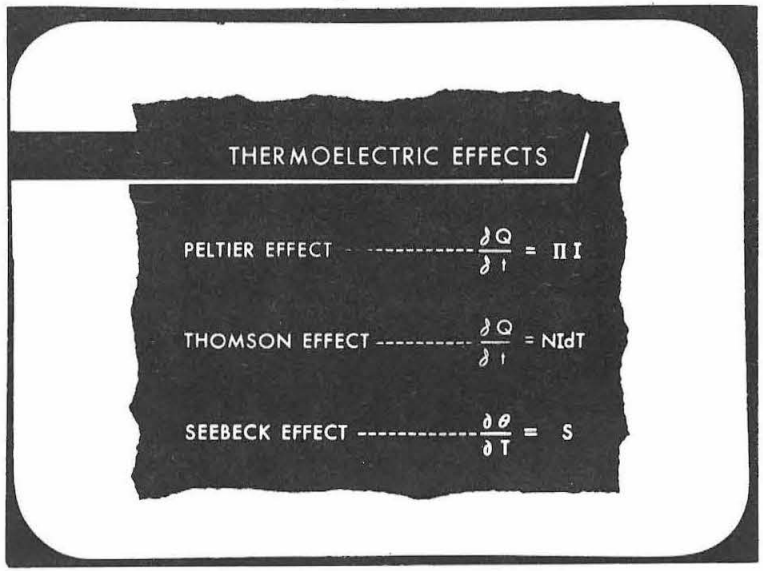
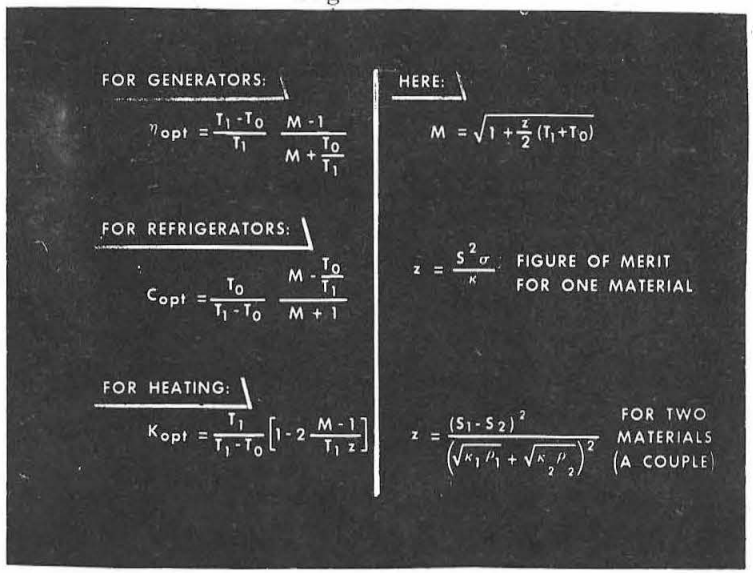


Figure 2



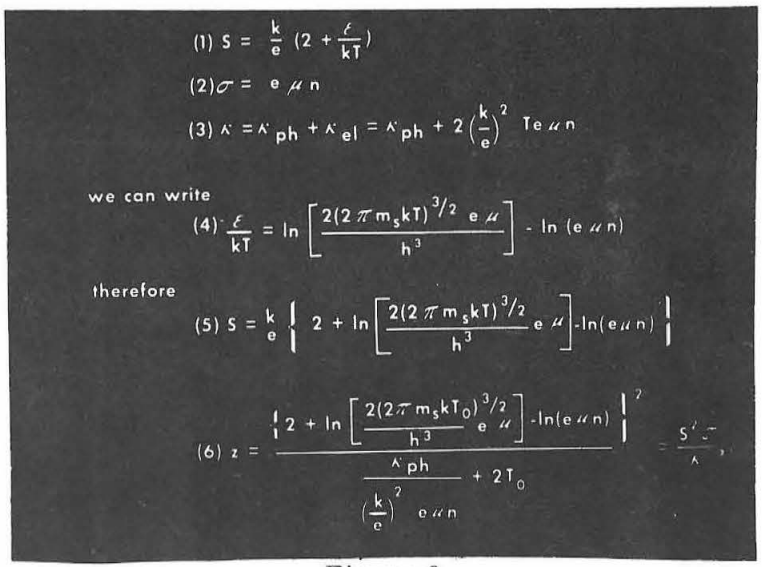


Figure 3

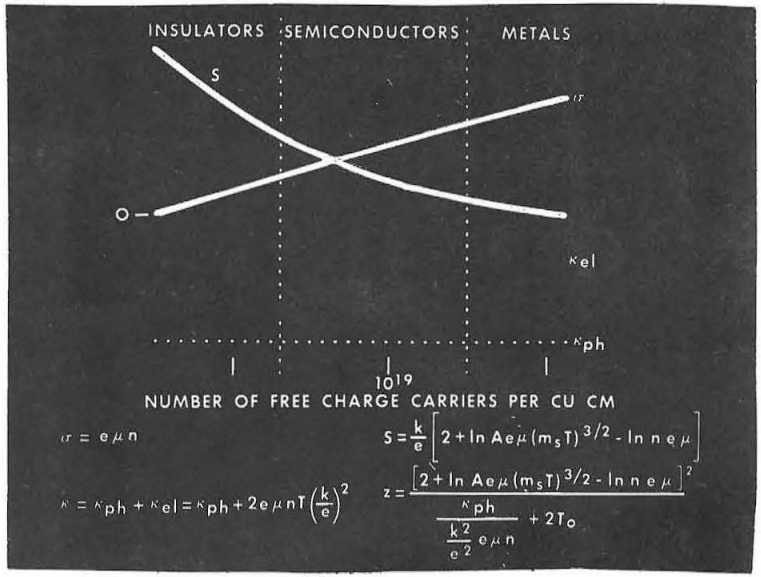


Figure 4

Figure 5

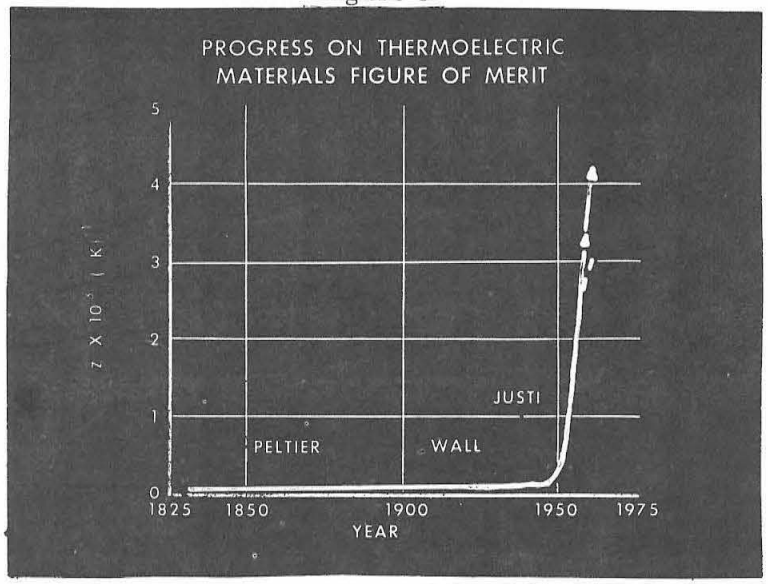
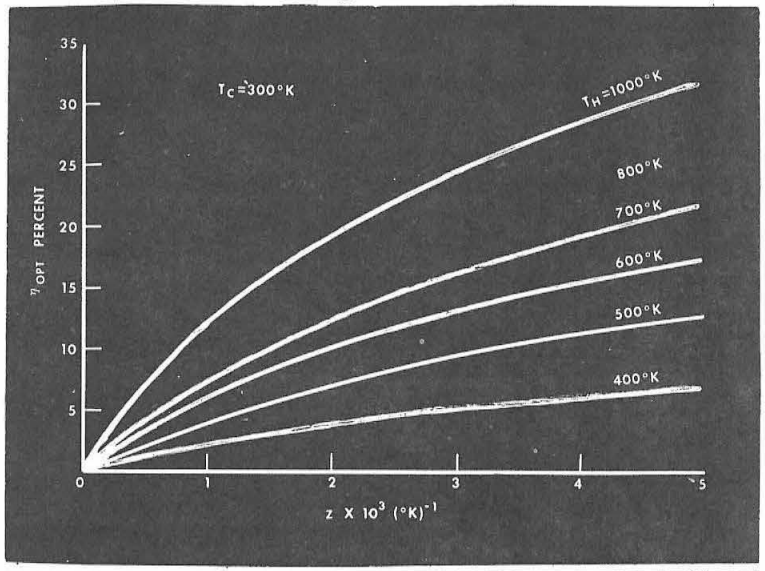


Figure 6



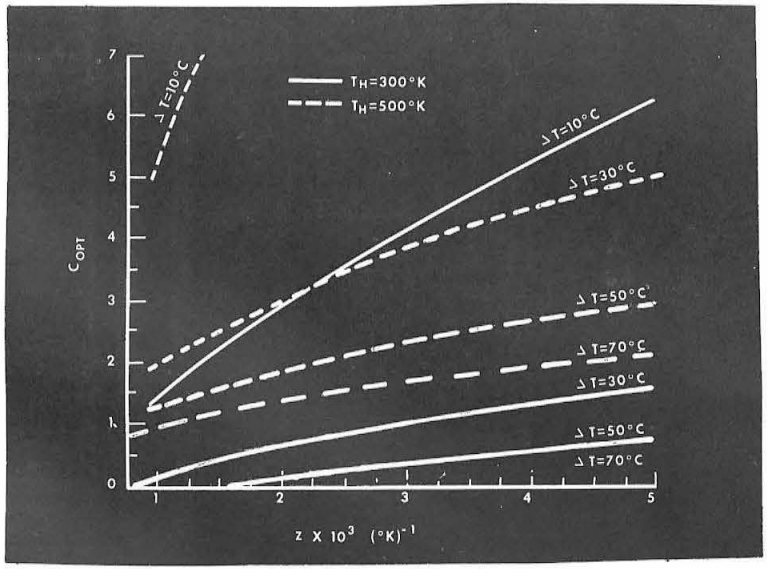


Figure 7

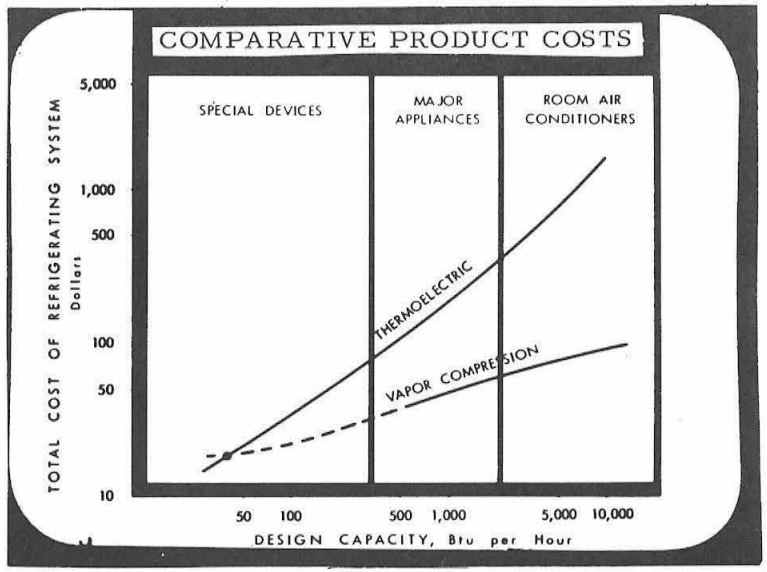


Figure 8

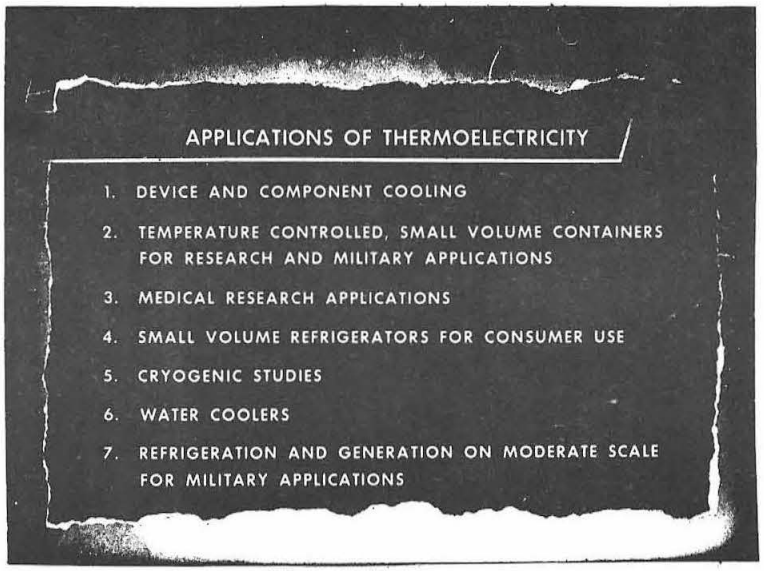


Figure 9

SYSTEM	TEMPERATURE	BEST RESULTS
I-VI COMPOUNDS Ag WITH Te AND Se	TO 600°C	Ag_2Te $z=1.3 \times 10^{-3}$ Ag_2Se $z=2.5 \times 10^{-3}$ $AgSbTe_2$ $z=1.9 \times 10^{-3}$
III-V COMPOUNDS Ga AND In WITH As AND Sb	700°C	In Ga As ALLOY $z \sim 1 \times 10^{-3}$
IV-VI COMPOUNDS Ge AND Pb WITH Se AND Te	PbTe TO 550°C GeTe TO 900°C	PbTe $z = 3 \times 10^{-3}$ GeTe $z = 1.2 \times 10^{-3}$
V-VI COMPOUNDS Bi AND Sb WITH Se AND Te	Bi_2Te_3 TO 350°C Sb_2Te_3 TO 350°C	$Bi_2Te_3 + Sb_2Te_3$ $z > 4 \times 10^{-3}$
I-III-VI COMPONENTS (CHALCOPYRITE STRUCTURE) Ag, Cu-Ga-Se, Te	~550°C	$CuGaTe_2$ $Z \sim 3 \times 10^{-3}$
RARE EARTH CHALCOGENIDES Ce, Sa, Gd, Th WITH O, S, Se, Te	1000°C	CERIUM SULFIDES (Ce_3S_4) $z \sim 1 \times 10^{-3}$

Figure 10

PHOTOELECTRONIC DEVICES

Albert Rose
Radio Corporation of America

In this discussion I should like to attempt to answer three questions on photoelectronic devices: (1) What are the goals of these devices? (2) What is the current status? (3) What are the directions in which we may try to make further progress toward achieving the goals?

In this field the goals are particularly easy to state. The purpose of a photoelectronic device is to see every photon that is incident upon the device. If the device can do this, then all the information which is incident on it can be gathered and used. I should like to explain what it means to see every photon that is incident on a device. The illustrations represent what a perfect device would see of an individual located about arm's length from the pick-up device, the device recording each photon that is incident upon it. This perfect device might be thought of as the human eye looking at the subject and recording each photon.

Figure 1 is what the device would see at a brightness near the absolute threshold of the human eye - about 10^{-7} foot lamberts. This is about 10 million times lower than the intensity of room light. What I particularly want to point out is that each photon is visible as a single speck on the illustration. These, by the way, are real photons. The device that was used to record this picture is a light spot scanner and therefore is not a normal pick-up device, but a highly useful device for simulating what the other devices would do. So these are traces of real, individual photons.

Figure 2 shows what one would see under conditions of a rather dark starlight night with cloud cover. The light intensity is on the order of 10^{-6} foot lamberts. At this point it can be seen that there is an individual, perhaps the sex is not confirmed, but merely the presence. Again it will be noticed that each photon is visible as an individual speck. It can be seen that the amplification of the device does not have unlimited virtue. Once the individual specks are seen, it does not make any difference how bright they are. The information transferred by a stream of light is limited by two things: one, that these are discreet specks from individual photons; and two, the fact that these photons do not arrive in any regular fashion; that is, they cannot be put where desired. They come randomly so that one gets a certain fluctuation in numbers per unit area, commonly referred to as "the noise of the photon stream."

Figure 3 shows the subject under starlight. Here, you begin to see the noise, not of individual photons, but the noise of clusters of photons. The individual photons are no longer decipherable, but the fluctuations from their clusters can be seen because of quite fundamental random processes.

Figure 4 brings the light intensity up to an overcast night in which there is about a quarter full moon. This would be about 10^{-4} foot lamberts. It will be observed that it is still limited by the fluctuations in the arrival of photons which give rise to a rough, grainy noise. Figure 5 shows the light intensity raised to what we would get on a clear night with one-quarter moon, say 10^{-3} foot lamberts.

Figure 6 represents a fairly "salable" quality which resulted under conditions of full moonlight (roughly 10^{-2} foot lamberts). A picture quality is seen that is beginning to be limited not by the photons coming in, but rather by the noise of the recording needle.

In this series of illustrations there is indication of the enormous range of light intensities over which people ask photoelectronic devices to operate. A device may operate at one range of light intensity and may give almost all the information in the light stream, but may not operate very well at other ranges. Therefore, the problem of defining the performance becomes one in which one must include the light intensity in which the device is to be effective. To repeat, at very low lights, one must be able to see the individual photons, whereas in the higher light ranges, one needs only to see the noise fluctuations of these photon streams and perhaps can tolerate more system noise than he could at low lights.

So much for the goals. We can now take a look at the status of a number of photoelectronic devices, particularly those devices used in recording pictures - television camera tubes, etc. Figure 7 shows schematically what the status is. The accuracy is only approximate but serves to bring us reasonably up to date. The range of brightness is plotted logarithmically on the abscissa, going from 10^{-7} foot lamberts up to about 100 foot lamberts, or about 9 powers of 10 of light intensity, and I have indicated roughly where one finds the various familiar key points of brightness. On the vertical scale is plotted, also logarithmically, the picture quality obtained at these various brightnesses.

If one had a perfect device, the plot of picture quality would rise linearly with the brightness available, that is, the total photon stream incident on the device. The heavy solid line (upper line) represents the level of quality one would get if he had such a device. Further, by way of comparison, sketched in (lower line) is the performance that we have already achieved by a familiar device, the human eye. Two things are striking about the human eye. One is the remarkably close approach that it makes to the performance of a perfect device; namely, it acts as if it were making use of one out of every ten incident photons. Secondly, this is true over a range of brightness from about 10^{-6} foot lamberts up to room light, then begins to drop slightly at higher intensities. The eye is acting like a counter which is making an accurate count of all the photons that strike the retina within a factor of 10. The eye is one of the better solid state devices we have in this field.

Indicated in Figure 8 are several of the devices that are reasonably familiar on the current scene. The "image orthicon" is a television camera tube; the "vidicon" is a television camera tube used currently for picking up and transmitting pictures. The intensifier "orthicon" is an adaptation of the "image orthicon," designed to carry its operation down to lower light intensities. Also, by way of orientation, photographic film is shown in the upper right-hand corner. It should be noted that several electronic devices that have been designed and fabricated also approach, as does the eye, within a factor of 10 (some of them perhaps slightly closer) the "perfect device." In the case of the "image orthicon," the reason for not going all the way toward a perfect device is that the photocathodes which these devices use deliver one electron for from 3 to 10 incident photons. If they delivered one electron for each photon, they would be ideal devices.

The vidicon is a camera tube which makes use of a photoconductive target. The image orthicon makes use of a photocathode intensifier as does the orthicon. The two solid state devices, the vidicon and photographic film, perform at rather high light intensities while the vacuum devices extend the range down to quite low light intensities. Photographic film is a very good solid state device, making use of electrons excited in silver bromide.

Figure 9 shows the comparative size of the image orthicon (upper) and the vidicon (lower), both television camera tubes. What I want to call attention to here, and to give a little more evidence of, is the advantage of solid state devices in this field over vacuum devices; they tend to be considerably more compact. The vidicon

shown is about the size of a piece of chalk. With this as a guide, if one thinks of doubling the size of the upper tube, he gets the intensifier image orthicon, the performance of which is sketched in an earlier illustration. If one thinks of half the size of the vidicon shown in Figure 9, he gets the size of the vidicon used recently in the Tiros satellite.

Figure 10 shows a picture that was transmitted by the half-inch vidicon from the Tiros satellite and depicts a part of the western bulge of North Africa. This turns out conveniently so that a square mile on the picture corresponds with one picture element of the tube. By a further elementary calculation one finds that the information contained in one picture element - that representing a square mile - corresponds to the addition of some 10^{-2} free electrons. Thus, the difference between zero light and full brightness in this picture element is 0.01 of a free electron per picture element.

Figure 11 is a schematic diagram of the intensifier orthicon which consists of a normal image orthicon with an electron image section tacked onto the end of it, so that electrons released from the photocathode will give rise to light at a phosphor-layer and that light then falls on the normal parts of the tube.

Photoconductors are one of the peculiar inversions of history. Even though they go back some 100 years and antedate the identification of the electron, it is only within the past decade that one can go out on the open market and purchase devices whose components depend upon photoconductors.

I want to discuss several of these devices. The familiar two-element photocell, the simple photoconductor, is small even by comparison with the vidicon. What is of particular interest is that in a simple two-element photoconductor cell, one can achieve a gain (that is the ratio of electrons to photons) of some 10^5 or 10^6 approaching a million fold. That is, one gets some 10^5 electrons for each photon that is incident on the device. He does not get them immediately; he must wait for them. The photon excites one electron - it circulates, and another electron comes in to take its place. Finally, after some 100,000 pass, the original excitation decays and completes the act. By rather elementary analysis, the performance of the simple photoconductor cell is inverse to the size and actually appears to diverge just as the device disappears; that is to say, as the electrode spacing approaches zero. By comparison, if one looks at the vacuum photo multiplier, it is more likely to be about the size of a fist.

Another small device goes by the name "light amplifier." A light amplifier has a layer of photoconductor and adjacent to it a layer of luminescent material; these two layers are laid on a support, a voltage applied on the two layers and by virtue of a potential divider action, the electroluminescent material lights up only where light is incident on the photoconductor. The light amplifier, by the way, is an exceedingly good example of the fact that photoconductors are really power amplifiers. For this device, if one puts in light on one side he finds that 100 times as much light emerges on the other side. In order to do that with a photoconductor and electroluminescent material, the photoconductor itself must exhibit a power gain of some 10^5. The reason for this is that the electroluminescent material has an efficiency of only 10^{-3}; then, to give the additional light amplification, one needs 10^5. This is also an example of an instance where, as in most of these devices, one is interested in the power gain features of the photoconductor and one is also interested in the product of power gain times speed of response, or the gain band width product - for photoconductors.

The last device I shall discuss is familiar to many of you, I am sure. It looks like an ordinary sheet of white paper; in fact, it is substantially just that. One extraordinary feature of it is that it is good quality paper because it has zinc oxide powder laid on the surface. This used to be done by paper manufacturers to produce a good white surface on the paper. If one makes the right choice of zinc oxide powder, then

this extraordinarily good looking sheet of paper becomes the basis for a simple photographic process. One takes such a sheet of paper, lays it on a metal plate, puts a charge on the surface, exposes it to a quick flash of light, then dusts the paper with most any sort of dust and in so doing, can obtain in a short time pictures that are of good contrast and of quite high resolution. This process is called "electrofax" and as you know, there is quite a successful parallel product to this manufactured by the Heloid Company under the name "zerographic process," which makes use of amorphous selenium. While I mention amorphous selenium and zinc oxide powder, it is perhaps worth reminding ourselves that not all the devices we want depend upon having nearly perfect single crystals. For some reason, there are a number of devices which appear to be quite successful commercially and which make use of some of the most wretched material one could imagine from the point of view of good physics or good analysis.

The vidicon shown in Figure 9 makes use of a photoconductor - antimony trisulphide which is evaporated on the front surface of the tube. It is evaporated through a poor vacuum so that a pile of submicroscopic sand ends up on the face of the tube. The particles are about 100 angstroms on the side and the material is very irregular from the point of view of physics.

The zinc oxide powder used in the electrofax process is also a powder of some 1000 angstroms on a side - 0.1 micron size powder. In fact, the key to its operation is the fact that it is about 0.1 micron, so that one can dope the outside surface of the powder with oxygen and be confident that the interior part of the particle is already doped by simply doping the surface.

As mentioned earlier, the zerographic process makes use of an amorphous material, amorphous selenium. If one makes the mistake of trying to crystallize the selenium into a nice crystal, one no longer has a device. Also, with regard to the zinc oxide - the electrofax process - it is worth pointing out that this material, even in its powdery form, acts as a metal semiconductor rectifier. That is, it retains the charge when charged negatively, but does not retain the charge when charged positively. In fact if charged positively, the charge would leak off in a small fraction of a second; but if charged negatively, the charge remains for some hours. The fact that the charge remains for some hours at a potential of some 500 volts - this is the back direction of the rectifier - means that this particular rectifier is passing something less than 10^{-10} amperes per square cm. with 500 volts back voltage, which is better than any of the semiconductor devices with which we are more familiar.

So much for the array of devices available. I should like to go back to the comparison between solid state devices and vacuum devices. Vacuum devices, as noted earlier, are successful in operation at very low light intensities. This stems from the fact that before excitation there is present in the device substantially no electrons; that is, the vacuum is a very good insulator. These devices are successful because one knows how to fabricate an electron multiplier for a vacuum operation and thus arbitrarily high gains can be had without sacrificing speed of response simply by putting in a number of stages of electron multiplications.

In the case of photoconductors, this is not true. Perhaps the gist of this message lies in the fact that in the case of photoconductors we have recently persuaded ourselves that the performance of a photoconductor is governed by a very simple relation; namely, if one writes down the gain - the gain meaning the number of electrons passed through the device for each photon - multiplied by the speed of response, then this so-called band width product is proportional to the conductivity of the photoconductor under its condition of use. That is, if one puts light on it, one has

conductivity. More specifically, I write the product:

$$\text{Gain} \times \frac{1}{t_o} = \frac{1}{t \text{ relaxation}}$$

Where t_o represents the time of response, i.e., the time it takes for photocurrent to rise or decay when the light intensity is changed; t relaxation in more familiar terms is the reciprocal of the RC product for the material with which one is working.

This is a remarkably general relationship, particularly because there is nothing here that says we are dealing with a photoconductor. We have measured gain, the measured speed of response, and the conductivity of the material. What is missing that one would normally expect to see is something having to do with the lifetime of free carriers, the mobility, the density of traps, and other features of the photoconductor - perhaps some geometry and the applied voltage. All of those are missing, and for that reason the relation is quite convenient and quite useful in guiding our expectations of photoconductors. In particular, if one wants to deal with an insulator (most of the devices which I showed you are insulators) then the relaxation time of these materials - or the 1/RC time - is on the order of one second. So, if one imagines on the right-hand side, a quantity unity; and at the time if one asks that the device have a gain or multiplication of say, 10^6, then the response time in this instance would be some 10^6 seconds. One would have to wait most of a day for the device to yield that amplification from a given pulse input.

In any of the devices one wants to use in the form of insulators - taking pictures, recording pictures - that amount of time is not allowed. In particular, the amount of time that is allowed is such that the response time turns out to be very close to the relaxation time of the material with a net result that the maximum gain, the number of electrons per photon that one can get in these devices, is unity. So, one finds in the vidicon that this is just what has happened. The best photo response of these devices is on the order of unity. The same is true of the photographic processes - electrofax and zerography. The reason that these results are obtained is because of the above relation.

One comment about the relation is that it has been derived by making use of the concept of space-charge-limited-current-flow in solids. That is the same sort of space charge limited flow in solids with which one is already familiar in a vacuum. Whereas space-charge-limited-current-flow in solids has been the subject of a certain amount of publicity as to its potential use in devices, I am inclined to be a little conservative about that aspect of it. On the other hand, its negative aspect, its aspect of limiting the performance one can obtain from photoconductors, is quite real, quite useful, and well attested to; and that is the aspect contained in this relationship.

Another aspect of the relationship is that while this relationship is true for most photoconductors, there are some very exceptional circumstances involving a rather sophisticated distribution of traps and recombination centers in the material where the performance of the device can be enhanced by a factor of m where m is greater than or equal to unity. That is, m is normally unity but can, if one is sufficiently skilled in the design of materials and in the distribution of defect states, exceed unity.

Another point worth mentioning is that if in the devices, say the photographic process, the electrofax or the zerographic process, or in the vidicon, one could depart from this stringent relation by a factor of 10, the economic consequences are easily measured in the millions. In the case of the photographic process, they would approach closer to the 10^9 figure. Then there would be a simple electrostatic process to compete with the ordinary chemical process. It is possible that one can get better performance from the photoconductors by the choice of recombination states and traps

leading to an m value somewhat greater than unity. Whereas this may very well take place, it is going to be a slow process technologically.

There is another avenue by which one could markedly exceed the limitations given by the above expression. As previously mentioned, if one wants a gain in the device of 10^5, excited one electron and then waited until 10^5 have passed through, it would be a long time. However, if in this device an electron has been excited and by impact ionization created two and they created four, etc., one would have a solid state amplifier within the simple piece of insulator. It would be an electron multiplier built in a solid precisely analogous to what one has in a vacuum tube. The phenomenon of impact ionization is well known; there are plenty of examples of it. That part is not novel, but the difficulty here in engineering such phenomena into a photoconductor is that one wants to be able to apply an electric field high enough so that if one electron is added, this electron is multiplied 100, 1000, or 10,000 times. However, no dark current should be present before one electron is added. The dark current normally is present by virtue of incidental effects such as tunneling from the metal electrode into the insulator or a zener emission from the valence band of the insulator into the conduction band. Therefore, the margin that one has to work with between where he can get photomultiplication under control and to where he runs into quite severe competition from other effects is very narrow and the problem exceedingly delicate. It is sufficiently worthwhile, however, so that I would think the future of photoconductors lies in getting this process under control.

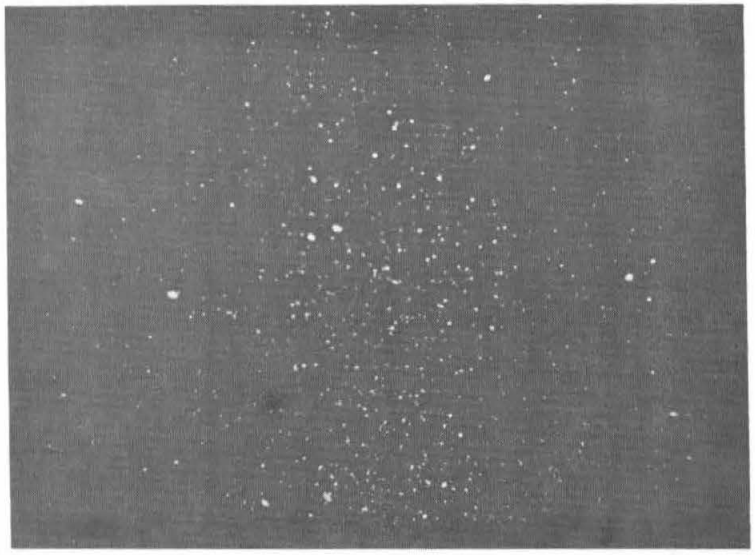


Figure 1

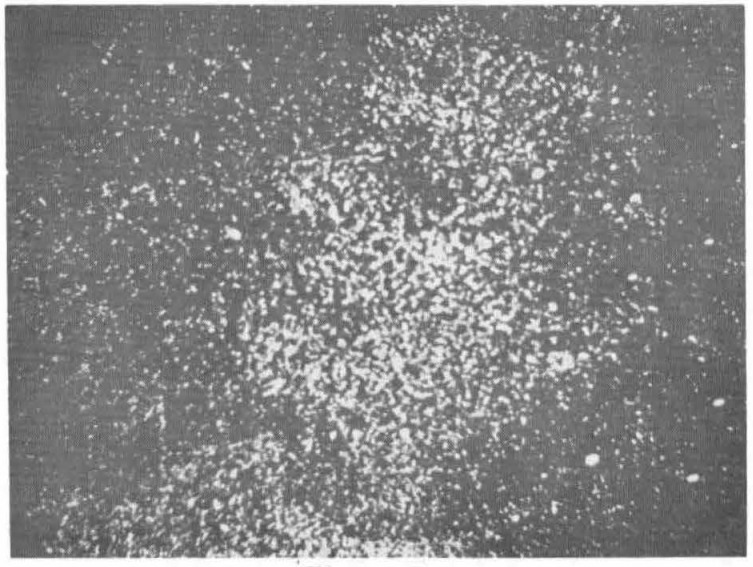


Figure 2

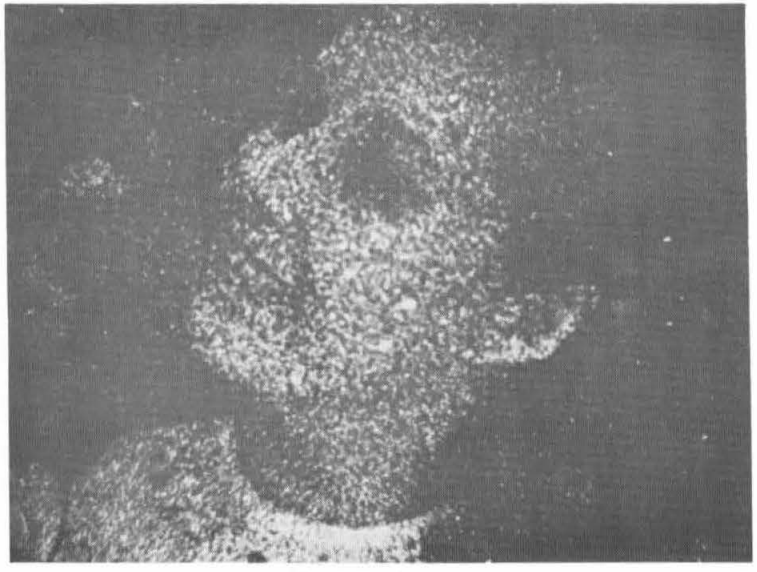


Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

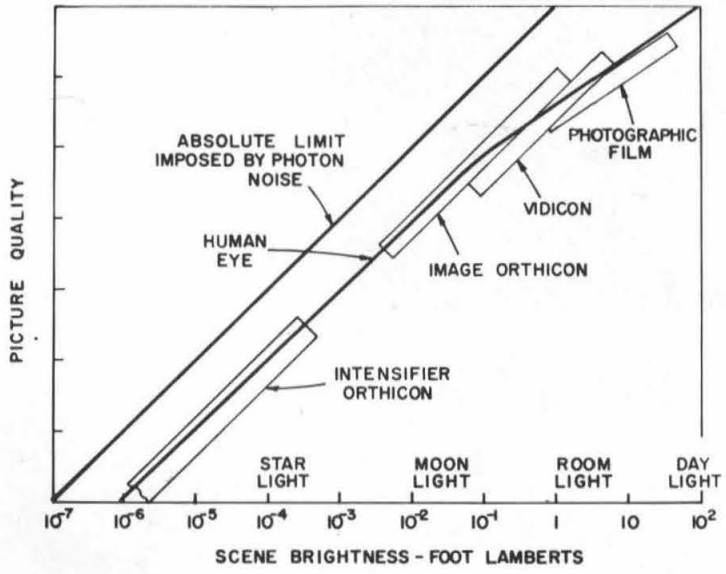


Figure 8

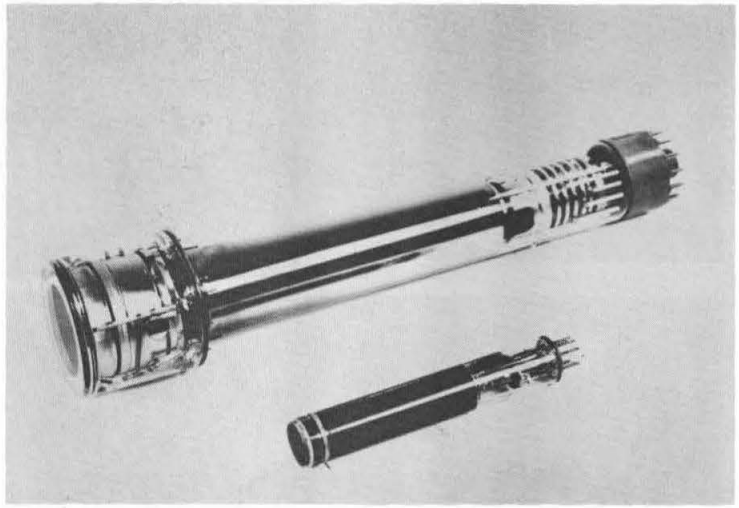


Figure 9

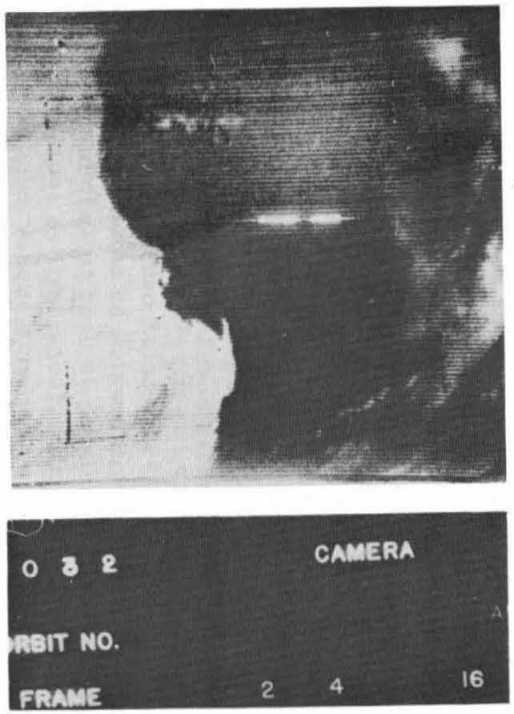
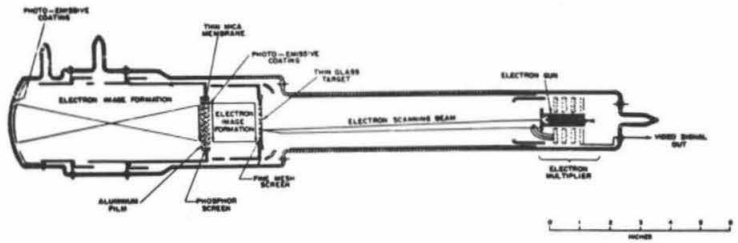


Figure 10



SINGLE STAGE INTENSIFIER ORTHICON

Figure 11

CRYOGENIC DEVICES

William B. Ittner, III
International Business Machines Corporation

In order to efficiently utilize high speed switching elements to form an information handling system, it is mandatory that the time required for information to propagate between the various elements be small. The ultimate speed of a very fast computing machine is thus intimately related to the over-all size of the system, independent of the particular switching elements used in its construction. With all solid state technologies, the extent to which an operating array can be miniaturized is generally determined, aside from the difficulties of fabrication, by the degree to which it is possible to extract the heat generated in its operation. In this regard, superconducting or cryogenic devices, because of their inherently low-power levels, offer considerable potential. These considerations, along with the potential low cost of cryogenic elements and the high reliability which is to be expected from elements operating in an inert low-temperature liquid helium environment, have provided considerable incentives to attempts to reduce the cryogenic technology to a practical and marketable state. The development of various cryogenic devices, their speed and power level, and a number of problems encountered in their development are reviewed.

Basically, a cryogenic computer is composed of a number of loops formed of two parallel superconducting paths, either of which can carry a current from a supply to a common ground. In the steady-state condition, the supercurrent is established in one of the two parallel paths where it may, for example, represent a binary zero or one. Because the direct current resistance of a superconductor is identically zero, there is no ohmic heating associated with the steady-state of a superconducting circuit; and, consequently, there is no dissipation of power. The circuit can be caused to change its state by the application of a control signal which momentarily destroys the superconductivity of a portion of the path carrying the supercurrent, and consequently forces the supply current to the alternate superconducting path. During the switching process, heat is produced by ohmic dissipation in the non-superconducting path, but as will be seen, the total power involved is relatively small. Thus, a superconductor has the ideal property, that it only dissipates power when it is actively involved in transferring information from one position to another. This paper is concerned chiefly with the characteristics of a number of devices which can be used to accomplish the switching process outlined above.

The present interest in cryogenics, as it relates to the computer industry, was sparked by Dudley Buck, who in 1955 pointed out many of the potentials inherent in the use of a superconducting switch which he named the "cryotron." It is the cryotron, which in one form or another is used to switch current from one superconducting branch to another. Basically, these devices consist of two superconducting elements, a "control" which always remains superconductive and serves to provide a magnetic field for switching a second superconductor, the "gate." It is the resistance of the gate when it has been driven into the normal state which drives the supply current out of one line and into another. The continual superconductivity of the control is assured by using a material which remains superconducting in magnetic fields and at temperatures sufficient to destroy superconductivity in the gate. In order that one cryotron be able to drive one or more additional cryotrons, it is necessary that the device have gain. In other words, the gate must be able to carry a larger current than is required by the control to drive additional gates. It is possible to provide for gain by the proper choice of geometry for the control

and the gate. The first cryotrons, for example, were simple wire-wound devices in which the gate consisted of a straight wire of moderately large diameter. The control, a much finer wire, was wound around, but insulated from, the gate in the form of a solenoid (Figure 1). Since the field produced by current flow in the control is larger than the field produced by an equal current flow in the gate, a small current in the control is able to switch resistance into a gate capable of carrying a large current. Thus the device has gain.

In a rigorous sense, the speed of a cryogenic circuit is determined completely through considerations of the characteristic impedances of its elements. In most practical cases of interest, however, the speed with which a current can be transferred from one path to another is determined by the ability of the driving force (developed as a potential across the resistive element of the cryotron) to overcome the inductive inertia of the two branches comprising the superconducting loop. The transfer time is thus determined by the inductive time constant of the loop which is just the ratio of the total inductance of the loop, L, to the resistance of the active cryotron gate, R.

The actual operation of a given loop proves, upon close examination, to be relatively complex, and involves a number of electromagnetic and thermodynamic processes. It has been found, however, that the actual "operating speed" is always close to the circuit L/R time constant, and the time constant is thus a simple and convenient parameter for describing circuit operation. In this paper we shall speak of approximate time constants for a single device. It must be remembered that the loop time constant is generally an order of magnitude greater than the individual device time constant.

It can be shown that while there are a number of heating processes, both reversible and irreversible, which take place during a switching cycle, the only important energy dissipation is that associated with transferring the current in a loop. The energy involved here is just the sum of the stored electromagnetic energy at the start and at the conclusion of the switching process. This, in turn, is just equal to $1/2LI^2$ where L is the loop inductance and I is the transferred current. The power associated with the switching process is given by the energy dissipated, $1/2LI^2$, divided by some multiple of the circuit time constant, L/R, and is proportional to I^2R. In the steady-state case where the device is run continually, the power dissipation is thus proportional to I^2R.

It is immediately obvious that the circuit time constant can be reduced either by reducing the circuit inductance, L, or by increasing the cryotron gate resistance, R. Decreasing the circuit inductance not only makes the circuit operate faster, but lowers the energy per switch, so that a higher speed is obtained at no increase in the steady-state power dissipation. Increasing the circuit resistance, however, increases the power dissipation.

In the original wire-wound cryotron shown in Figure 1, the inductance could not be reduced much below a microhenry, and the resistance could not be conveniently raised above a milliohm. The L/R time constant for a single device is, consequently, of the order of a millisecond. Because of its slow speed, the wire-wound cryotron was replaced almost immediately by the thin-film cryotron, which possessed two distinct advantages. It is obvious that the small cross section of a thin film results in a substantial increase in cryotron resistance, but more important, by using the diamagnetic properties of a superconductor, (namely, its ability to confine magnetic fields) it is possible with thin films to substantially lower the inductance of the structure. This is achieved by forming the thin-film cryotron structure (Figure 2) over, but insulated from, a superconducting base plate (called the ground plane). The films comprizing the cryotron itself, along with the superconducting ground plane, form, in effect, superconducting transmission lines, wherein the magnetic field

generated by current flow in the films is confined to the region between the film and the ground plane. The presence of a ground plane results in a substantial decrease in the over-all inductance of the thin-film cryotron structure. Typically, the inductance of a crossed-film cryotron constructed from metal and insulating films 5000 angstroms thick is of the order of 10 micromicrohenries. The resistance of the gate film is generally of the order of 10^{-3} ohms. Hence, the time constant for an individual cryotron is of the order of 10 nanoseconds.

It can be shown that the time constants of thin-film cryotron structures are independent of their planar dimensions. In principle, it would seem possible to obtain a further decrease in the time constant by simply decreasing the thickness of the structure. Decreasing the insulation thickness, for example, has the effect of decreasing the inductance of the transmission line, while decreasing the thickness of the metal film results in a higher resistance in the normal state. Here, however, a fundamental limitation is encountered in the fact that while a superconductor is almost completely diamagnetic, it is not entirely so. In fact, magnetic fields do penetrate superconducting films to a depth of about 500 angstroms, the so-called superconducting penetration depth. This finite penetration, while small, has a profound effect upon the electromagnetic properties of thin films, and in particular, it almost completely determines their behavior as devices. The fact, for example, that the magnetic field penetrates the ground plane means that the ground plane is not completely effective in confining the magnetic field of the cryotron, and hence, the device inductance is higher than would be the case if the ground plane were a perfect shield. Thus, if the insulation thickness is decreased to reduce the device inductance, a point of diminishing returns is reached when the insulation thickness approaches about 2000 angstroms. Further decreases in the insulation thickness have little effect on the time constant of the device.

The finite penetration of a magnetic field in a superconductor has an even more profound effect on the electromagnetic characteristics of the cryotron itself, and is so intimately related to the basic nature of superconductivity, that it is worthy of some discussion. Fundamentally, the phenomenon of superconductivity arises because of the long range electronic interactions which occur in certain metals at low temperature. In all metals, electrons are thought to interact directly with one another through the electrostatic coulomb forces they exert on one another, and to interact indirectly with one another through mutual interactions with the vibration of the crystal lattice. In the latter case, an electron interacts with a lattice vibration (phonon) which, in turn, interacts with another electron. The effect, as far as the electrons are concerned, is equivalent to their having interacted directly with one another. The coulomb interaction is a repulsive one, and normally much larger than the interactions between electrons resulting from electron phonon interactions. In a superconductor, however, the interaction between electrons brought about by their mutual interactions with phonons is believed to be larger than the coulomb interaction and is, in fact, attractive. Thus, the net interaction is, on the average, an attractive one, and electrons are coupled together in pairs where they behave like Bosons. The condensation of normal electrons into the superconducting state can be thought of as a kind of condensation which takes place in a Bose gas at low temperature.

Because of the net attractive interaction between electrons, they tend to operate in a collective mode, and it is this cooperation which results in the supercurrents that flow unimpeded through a superconductor. The mean spatial range over which electrons interact with one another in a superconductor is believed to be of the order of 10^{-4} centimeters (called the superconducting coherence length). It is, therefore, not surprising that the superconducting behavior of a thin film whose dimensions are less than about 10^{-4} centimeters, should be quite different from the behavior of a bulk superconductor, and experimentally this is found to be the case. It might

be said that in general, when supercurrents are confined to flow in structures whose physical dimensions are less than the superconducting coherence length, the normal long-range interactions are diminished, and the specimen tends to lose some of its superconductivity, e.g., it tends to be less diamagnetic, and magnetic field penetration increases. This behavior of superconductors results in a thin film being able to carry less current per unit area than a bulk material, and thus a gate film which is thin loses its ability to carry the supercurrent necessary to operate the controls of successive cryotrons. In effect, cryotron gain is drastically reduced when the gate films are made much thinner than the superconducting coherence length.

The same sort of behavior is produced by adding impurities to a superconductor to increase its normal state resistivity, although the effect is less dramatic than that produced by size restrictions. Here again, the scattering of normal electrons (which are always present even though in the superstate they are short circuited by the super electrons) by the impurity atoms, tends to diminish the long-range interactions present in a pure bulk specimen.

From a device standpoint, the effect of increasing the gate resistance, either by decreasing the gate thickness or increasing the gate resistivity, is to reduce the cryotron gain. There are, therefore, limits to the speed increases which can be obtained by adjusting dimensions or by using high resistivity materials. It is possible, however, to obtain a considerable improvement over the crossed-film cryotron by using the technique of biasing.

In the cross-film cryotron of Figure 2, gain is obtained by making the control film narrower than the gate film, thereby obtaining a high field from current flow in the control. The necessity to restrict the width of the control adversely affects the cryotron time constant in two ways. First, since the magnetic field of the control produces resistance only in the gate which is directly under the control, a narrow control produces resistance in a correspondingly short length of gate. Secondly, since the inductance of the control is inversely proportional to its width, a narrow control has a high inductance. It would obviously be desirable to make the control as wide as the gate if it were not for the fact that the gain would then be less than unity. This difficulty can be resolved by the addition of a second control film which serves to bias the control to a level where only a small current is sufficient to drive the gate film normal (see Figure 3). The incremental gain of a unity ratio crossed-film cryotron can be quite large. By widening the control film the inductance of the structure is reduced and typically is of the order of 10^{-12} henries. Correspondingly, the resistance of the gate is increased to about 10^{-2} ohms. The L/R time constant for a single cryotron is thus in the vicinity of 10^{-10} seconds.

The biased structure has a number of potential advantages not inherent in an unbiased structure. For example, the bias control can be used either to provide a static bias, or can in fact, be used as a second control. In the latter case the cryotron can be used as a logical element that is switched by the presence of signals in both controls, but is not effected if only one of the two signals is present. The use of two controls, or one control and a bias, also means that each control will carry a smaller current than the single control of an unbiased cryotron. Since the power dissipation is proportional to I^2R, the power dissipation can be substantially reduced.

Another type of cryotron suggested by the concept of biasing, is the so-called "in-line" cryotron which is shown in Figure 4. In this configuration the control and gate are placed in line with one another, and gain, if desired, is obtained by the use of a second control or bias line. Since the L/R time constant of the in-line cryotron is independent of length, (i.e., both L and R increase directly with length) the time constant is of the same order as that of the unity ratio crossed-film cryotron. However,

the in-line type of cryotron has, in certain specific applications, advantages over the biased unity ratio cryotron. It can, for example, be used in a loop that must of necessity be long. Such loops are required both for coupling between substrates in the helium bath, and for coupling between substrates at helium temperature and the outside world. In these applications it is expedient to match the in-line cryotron resistance to the characteristic impedance of the line. When the cryotron gate is switched to the resistive state, a pulse propagates along the line with the characteristic velocity of the line. The price that is paid in using long in-line cryotrons is chiefly the increased power dissipation resulting from the increased resistance.

In summary, it is possible to build simple cryotron loops in which the loop time constant is in the vicinity of a nanosecond. It is also possible to build fast computing circuits which are composed entirely of short loops. In such circuits it can be shown that the delay per stage of signals propagating through a chain of loops is of the order of the circuit L/R time constant. The reset or recycle time is found to be about 3 or 4 time constants in length. Thus, in a selector switch, for example, it would be possible to start the second access to the switch after a signal had propagated through about three or four levels of the switch. While it is possible, in principle, to build circuits in which the time constants are less than a nanosecond, the ability to extract heat from such circuits limits the recycle time, i.e., the frequency with which they can be operated. It is ultimately a thermal process which sets the upper limit on usable speed.

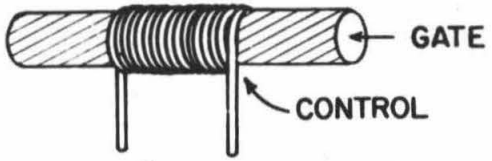
It would be inappropriate to discuss the thermal problem in any detail here, but it is interesting to emphasize the fact that the steady-state power dissipation can be shown to scale directly with the planar dimensions of the circuits. Hence, there are substantial gains to be attained in the microminiaturization of cryogenic circuits since the problem of power dissipation will not increase if circuits are compacted simply by reducing their dimensions.

At the present time we are attempting to build a number of high speed computer circuits to demonstrate some of the principles discussed here. Loops with nanosecond time constants have been built, and their calculated time constants have been found to agree with the measured values. A small prototype memory and a simple subtractor circuit have been built using unbiased crossed-film cryotrons. While these circuits were not expected to be fast, i.e., their operate times were of the order of microseconds, they did provide a means of verifying some of the theories of their operation, and simultaneously, provided a check point on the technologies involved in their fabrication. We expect this year to demonstrate a few simple prototype systems using biased cryotrons operating in high speed mode.

At the present time it seems probable that the first commercial applications of cryogenics can be expected in about 1964. We tend to look at cryogenic devices as one more tool in the arsenal of solid state devices. For many applications their mode of operation is ideal, and we are confident that cryogenic technology can make a substantial contribution to the ever growing array of solid state devices.

Figure 1

WIRE-WOUND CRYOTRON



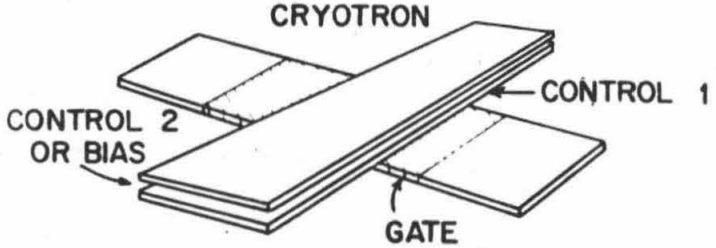
$L \cong 10^{-6}$ HENRIES, $R \cong 10^{-3}$ OHMS

$L/R \cong 10^{-3}$ SEC.

$1/2 L I^2 \cong 10^{-7}$ JOULES

$I^2 R \cong 10^{-3}$ WATTS

CROSSED-FILM BIASED CRYOTRON



$L \cong 10^{-12}$ HENRIES, $R \cong 10^{-2}$ OHMS

$L/R \cong 10^{-10}$ SEC

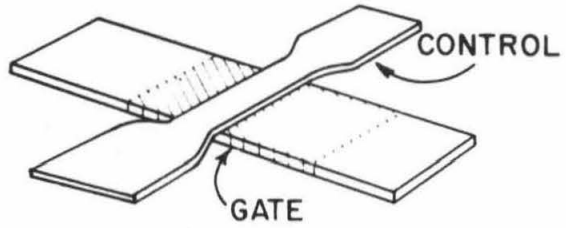
$1/2 L I^2 \cong 10^{-14}$ TO 10^{-15} JOULES

$I^2 R \cong 10^{-4}$ TO 10^{-5} WATTS

Figure 3

Figure 2

CROSSED-FILM CRYOTRON



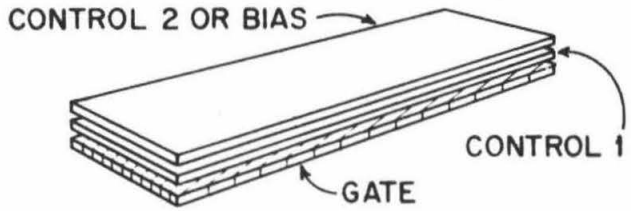
$L \cong 10^{-11}$ HENRIES, $R \cong 10^{-3}$ OHMS

$L/R \cong 10^{-8}$ SEC

$1/2 L I^2 \cong 10^{-12}$ JOULES

$I^2 R \cong 10^{-4}$ WATTS

BIASED "IN-LINE" CRYOTRON



$L \cong 10^{-12} \frac{\ell}{\omega}$ HENRIES, $R \cong 10^{-2} \frac{\ell}{\omega}$ OHMS

$L/R \cong 10^{-10}$ SEC

$1/2 L I^2 \cong (10^{-14}$ TO $10^{-15}) \ell/\omega$ JOULES

$I^2 R \cong (10^{-4}$ TO $10^{-5}) \ell/\omega$ WATTS

Figure 4

THE TECHNOLOGY OF SEMICONDUCTOR DEVICES

John W. Peterson
Pacific Semiconductors, Inc.

My task is to tell you something about how one makes semiconductor devices. It is perhaps worth bearing in mind that not only are they an important class of device in their own right, but one might consider them the forebearers of a broad continuing line of additional types of new and very useful solid state devices that goes on as far as one can imagine into the future.

In order to understand why we do what we do in attempting to make these devices, it is probably worthwhile trying to describe what we are trying to accomplish, and why. In this connection it is helpful to think of a semiconductor device as being very much like an electron tube, in the sense that its action, in almost all cases, is a result of a movement of positive or negative charges. In distinction with electron tubes, we have mobile positive charges as well as negative charges, whose motion is controlled by the equivalent of electrodes. So, in making such a device, the objective is to place electrodes correctly, of the proper dimensions and spacing, and to do it in such a fashion that the moving charges can move under optimum conditions.

In comparing a semiconductor device with a tube device, there are certain things that should be kept in mind. One is that the velocity of the moving charges is relatively low (of the order of 10^6 to 10^7 cm/sec) in a semiconductor compared with the velocity that free electrons can have. The velocity of free charges in a vacuum tube can be as high as 10^9 cm/sec. This means that if we are going to get comparable frequency response, we have to have very close spacing in order that the time of transit shall be roughly the same in a semiconductor device as in a tube. Also, semiconductor materials have high dielectric constants and this means that if we are to have adequately low capacitance, the area must be made very small. Thus, a number of influences combine to require that a semiconductor device be very small and with minimum thickness. It turns out that the ultimate size for a given device is set by the thermal dissipation requirements because we are fortunate in that the semiconductor material can tolerate a rather high current density. Current densities of a thousand or more amps/cm^2 are quite practical and realized in most high performance devices. Therefore, nature has been kind to us, and I suppose this is why the industry has grown. We require small sizes, but on the other hand, the small sizes are electrically possible. Therefore, heat dissipation typically is the ultimate determinant of the size of the device, and usually because of this the device is a little bit bigger than we would like to make it in order to have optimum performance.

Another key fact, I think, which perhaps more than any other explains the exponential explosive growth of our industry, is the extremely high designability. I think anyone who has looked into the theory of semiconductors cannot help but be impressed with the rather naive assumptions that we make, and by the fact that these extremely naive assumptions actually work so well that we can forget all about the quantum mechanics we could go through to justify their use. By simply considering the small particles in the device that are negatively and positively charged and that are free to move, and designing the device accordingly, it will usually work about as one wants it to. This very simple, adequate theory of the semiconductor makes it easily possible to design something that one wants, and if it does not work right the first time, there is a fighting chance of determining why. On the other hand, perhaps the contact transistor suffered seriously because its theory was not well enough understood so that it could really be significantly improved by design theory. I am referring to junction devices, and the point contact transistor does not truly fall in this category.

The paper by Shockley in 1949 which described the PN junction laid the groundwork for the whole industry, and it is remarkable how nearly complete a treatment he gave of the properties of PN junction.

The desirable features of the semiconductor, its ruggedness, etc., that come because it is a solid, and the fact that the atoms are close together, puts rather extreme requirements on structural perfection and on purity. We take it for granted that we have a single crystal material when we start. Perhaps many of you might not even have seen a single crystal - they are not common. We have to have much more than a single crystal, though. We have to have a very low dislocation density, and a degree of perfection which is really rather ridiculous, except that we need it.

The purity has to be approximately one part out of a billion. Everybody spends their time thinking of ways to drive this number home. I do not know whether this one serves the purpose. If one were to spend whatever time was required walking up and down all the miles of railroad track - and there are over 200 thousand of them in this country - and if there was one bad tie - that is about the proportion of impurities that we can tolerate in semiconductor material.

In making a device we start out with the pure single crystal of very high perfection, and then we produce the semiconductor structure in it. That is to say, we make the electrodes. We do this by what we call doping. As Dr. Dacey described yesterday, if one wants an excess of free negative carriers, one makes N-type silicon. This is done by putting in an element from the fifth column of the periodic table - phosphorus arsenic. If one wishes a region which has free positive charges (P-type), an element should be put in from the third column, such as boron or aluminum. In doping one must consider the lateral extent of these "electrodes." Finally, contacts should be made to the various regions and the whole packaged in such a way that it will continue to be useful. This has been a long, hard battle, but we are making a great deal of progress now in producing devices which are beginning to deliver the long-term reliable operation which we have every right to expect should come from something which is only made of a piece of solid.

In general the hard materials today are silicon and germanium. There is considerable interest in intermetallics, but they have not had a great deal of commercial impact yet. Whereas the speaker on ferroelectrics made a plea for fewer new exotic materials and more good ones, we are in the very fortunate position of having relatively very few materials, all of them very good. Silicon is a remarkably fine material to work with, and it is coming more and more into its own as we improve our technology. Silicon offers high-temperature operation with junction temperatures to perhaps 200°C, and high inverse impedance, if that is important. Germanium offers somewhat lower temperature operation but has the advantage that high frequency response is easier to obtain because of the higher charge mobility. Also, germanium requires less forward voltage for a given current. Gallium arsenite, for example, offers high mobility and still higher operating temperatures and might ultimately take over some of the applications of silicon.

To start producing a material one must first refine the raw product. In the case of silicon, one starts with clean quartz sand. In one process the sand is reacted with zinc chloride to yield trichlorosilane, $SiHCL_3$, which is then reduced in the presence of hydrogen to produce a very pure silicon. Ironically, in the case of silicon, it seems easier to start with clean sand than to take scrap silicon and repurify it. Silicon was perhaps the first of the modern semiconductor materials to gain considerable attention and yet it arrived in practice later than germanium because the technology of producing pure, perfect crystals of silicon was so much harder than for germanium. Because of silicon's much higher melting point, germanium got there first.

Germanium is found in association with coal. In England, I believe, they use soot as a source. In this country it is obtained as a by-product in the refinement of other metals (zinc, lead). It is obtained as the dioxide which is reduced in the presence of hydrogen.

Some think of silicon as relatively expensive and germanium as relatively inexpensive because this is the way the prices of the devices have been. Actually, the story is turned around for the raw material. Silicon is cheaper than germanium, and because it is so very plentiful, one can expect the price of the crystals of silicon to go down a considerable amount.

The first successful method of producing single crystals for use in our industry was the Czachralski process. In this process one starts with a crucible of molten material, dips a single crystal seed into it, and then slowly pulls it out to grow an icicle of the solid material. There is a certain amount of purification in this process, but the result is rather nonuniform in impurity distribution. The impurities tend to segregate toward the bottom end of the ingot, because, in general, impurities would prefer to be in the melt than in a crystal. The concentration of impurity atoms in the melt increases, and the concentration of impurities in the crystal also increases so that the latter part of the crystal is more heavily doped than the first part. The term impurities perhaps has a bad connotation, but that is what they are.

To make a device one should have a material of a certain resistivity - a certain level of donors or acceptors. These impurities permit the electrons to move, and the most convenient way to measure their concentration (this is a tremendously powerful tool for us) is by measuring the resistivity of the sample. It is worth pointing out that measuring the resistivity makes possible the measuring of the donors or acceptors (impurities) to one part per billion. When our industry started, it was sheer luck if one part per million of impurities could be measured; so, detectability has been improved by three orders of magnitude.

Figure 1 is an illustration of the zone leveling process which is very successfully used for germanium, but not for silicon. In this method little molten regions are produced in an ingot from germanium which lies in the boat, and molten zones are swept through the ingot. The ingot and boat are pulled while the heating coils remain fixed so the molten regions move along the ingot. In effect they pick up the impurities and sweep them out of the ingot toward the tail end. In this fashion the detectable impurity in the material can be decreased to one part in 10^{10} which is quite remarkable. This method, unfortunately, does not work on silicon because silicon is an extremely reactive material. Its melting point is 1400°C, and it will react with the crucible, leeching impurities out of the quartz, which is the most impure element of this system. One of the impurities in quartz is boron, which has the unfortunate property of not preferring to remain in the melt as most impurities do. So the sweeping process does not work well in that case. Also, the silicon wets the quartz, and it is apt to break. To get around these problems the floating zone process was invented.

In the floating zone process (Figure 2) we do the same thing, except there is no contact between the molten region and the container. The polycrystalline rod is held vertically, and the molten zone is formed in it by the RF coil. The molten zone is first formed at the bottom of the rod where it touches a seed crystal and then gradually moves upward through the ingot, thus sweeping out all of the impurities except boron and leaving behind a single crystal of very pure and very uniform silicon. We get rid of the boron by eliminating it beforehand. There is really no other solution. The object in refining silicon is to go through a process in which it is very easy to remove the impurity compounds and leave the silicon compounds highly pure. The floating zone process has become quite commercially

important. It, unfortunately, does not offer the same large diameter that the crystal-pulling method does, but there is much effort in trying to improve this.

Figure 3 is an illustration of the kinds of ingots with which we work on a routine basis. The one on the bottom is a silicon ingot. The color comes from a silicon oxide produced by reaction with the quartz crucible. The one on the top was a floating zone refined. It is almost drill rod size. A few years ago we felt very lucky if we had a single crystal, even if it were turnip-shaped or hour-glass shaped. The crystal on the bottom of Figure 3 is an ingot of unusually high perfection, grown by very careful techniques and with a very small seed, and may even have zero detectable dislocations in it. This does not mean that there are none, but it does mean that we were not able to find any.

After an ingot is grown with as uniform dimensions and resistivity as possible, it is shaped by slicing, lapping, and etching to remove the lab damage, to give the slices which are typically what is used to make the devices. The junction forming may be done on a whole slice or it may be done on small pieces.

Each method that we use in producing the structures or junctions has its particular virtues and lack of virtues, and one picks a process which fits what one is trying to make. There are many processes and variants, and I shall only discuss a few. Also, I am going to discuss processes largely in terms of the transistor, because in general, this is the most demanding structure there is, and the various processes can be described very easily in connection with it. The same processes are used for many other sorts of devices as well.

The requirements of a transistor that is of high-frequency capability are low capacitance, low resistance, high power, high-voltage ability, and high-current gain. All would be ideal but of course this cannot be because there are always compromises in a typical engineering problem. The device itself is a compromise. However, one would like to attain a structure which reduces to a minimum the number of compromises that are necessary. We are always on the lookout for a process which will break to some degree the constraints of having compromises where if one property is improved, something else will be worse.

Figure 4 shows the desirable features of a more or less idealized transistor. There are other structures that are better for specific purposes but Figure 4 illustrates the basic organization. To obtain high-frequency response, a transistor should have a very thin base layer, and the collector region should be only as thick as is actually required to support the applied voltage. Also, the resistance should be low in the emitter and collector contacts. In Figure 4 these regions are very thin: the emitter is perhaps a micron in thickness, the base perhaps a micron in thickness, the collector from two to five microns in thickness and the collector contact will be a few mils. The collector contact can be thought of as a low-resistance handle since the collector base and emitter are too thin to handle, and it is necessary to have something to hold them during the assembly of the device. The metal contacts should be as close together as possible to reduce the spreading resistance.

While we desire thin regions and low resistivity for high-frequency response, for high-voltage ability we desire high resistivity and thick regions because the voltage is supported by the so-called depletion layers, and we do not want to limit the width which this layer can achieve.

In order to keep the capacitance down, the emitter and collector-base areas should be as small as possible; and because the current tends to flow at the edge of the emitter, the central portion is rather useless, but does contribute to capacitance. However, for high-powered ability the areas must be large in direct contradiction for the requirement for low capacitance.

Figure 5 shows a schematic of the grown junction process which can be applied either to diodes or transistors. In this process a crucible full of molten silicon is doped N-type by the addition, for example, of phosphorus. A single crystal is grown out part way and then a pellet of P-type, boron, dope is dropped into the melt which converts it to P-type. A thin layer is grown and then another pellet of N-type is dropped into the melt which converts it back to N-type and more ingot or crystal is grown. This gives a transistor or a potential transistor. The ingot is cut up into little bars and finally one has the final structure shown.

This is the first method that was used to produce actual junctions, and it worked remarkably well. However, it has a considerable number of disadvantages. It leads to a very thick emitter and collector region, which results in high resistance. It is very difficult to make a low-resistance base contact to it because one does not have access to the planned view of the transistor that occurs in the idealized structure. Also, it turns out to have an interdependence between the doping levels in the emitter base and collector which limits the attainable structures.

The grown junction process was followed by the alloy junction process which was a great improvement in many respects and is still very widely used today. Figure 6 illustrates the formation of a PNP germanium transistor by this process. It is used on germanium and silicon NPN and PNP-type transistors. In the case shown a small piece of indium is pressed against a piece of germanium and heated in a furnace to a moderate temperature - perhaps 5 or 600 degrees Centigrade. The indium dissolves the germanium on both sides as shown. When it is cooled off, the germanium precipitates back out of the indium, but it carries some indium into the lattice with it. Thus, one can achieve doping of the regrowth regions. It can also be carried out by using something like lead, which is not in itself a doping material, but one can put arsenic, for example, into it which will be left behind in the lattice. The lead which is left behind in the lattice apparently has no effect.

The disadvantages of this process are that it produces a very thick base region because the indexing in this process is done from the two surfaces of the wafer, and it is most difficult to maintain close tolerance. One is doing very well if the tolerance is maintained to within a tenth of a mil, which is not close enough control of the base layer thickness for the high-frequency transistors that are often needed. Also, the base resistivity determines the collector base voltage which is a limitation in design. The alloy junction process however, is very simple and economical and gives very low-resistance contacts so that for certain applications it is advantageous to use it, particularly where high-frequency response is not a problem.

Figure 7 illustrates a process developed by the Philco Corporation which produces what is called the surface barrier transistor. It first involves electrolytic etching, where jets of a solution are directed against a slab of semiconductor. By a very clever technique they can regulate and determine the thickness of the remaining web quite accurately. Light is shone down the jets until the light starts being transmitted through the web. Then the polarity can be reversed to plate metal on the two regions that were etched. The finished transistor from this process is shown.

The surface barrier transistor is an interesting example of a combination of processes. After the plating step the device is heated in a furnace which causes an alloying at the junction. This is called microalloying in this case - which greatly improves the quality of the junction that was formed. The disadvantage is that the device has a low-collector voltage if it also has good base resistance. One has a limited possible geometry here because again one does not have access to the planned view of the device. But, for its purpose it is a good process.

Figure 8 is the schematic of a diffused base transistor, and illustrates the solid state diffusion process. This process was pioneered at Bell Laboratories as were so many of the keystones of our field. It is the most widely used process that is in large scale use today. Because of its very high degree of versatility, it can be used to make almost any kind of device required. It is not necessarily true that it is the best process for a given device. However, in most cases it is, and in many cases, such as very high frequency or very high-powered transistors, they cannot be made by any other process. The solid state diffusion process is started with a slab of material which will form the collector region and diffuse in the base layer. This is done at high temperatures - perhaps 1200°C - in an atmosphere of the doping impurity required. For example, if this were to be an NPN, one would start with N-type material and diffuse in, for example, boron, to produce a P-type layer. A diffused region is produced on both sides of the wafer but one side is lapped off, so it is not shown in the diagram. The emitter region can be formed either by alloying or diffusion, depending on your choice. If it is done by diffusion, it is necessary to mask the surface so that diffusion will only occur in the region desired. This can be done by the use of silicon oxide which can be grown by using an oxidizing atmosphere at these high temperatures. The emitter (the white region at the top of the device) can then be produced by diffusing in, for example, phosphorus, which might have been the material with which the starting wafer was doped. I should make the point that in order to produce the emitter region we have to compensate for the doping in the base layer (hatched area) and overcome it. That is true in all of the other methods too. The black area at the top of the drawing to which lead b is connected is simply a contact to the base region.

The schematic of the reactor and process used in the vapor deposition or epitaxial growth process is shown in Figure 9. In this process the compound introduced into the reactor decomposes to deposit the desired material as a layer on the surface of the ingot in the quartz boat. This process has the advantage that it is taking place in a very clean system. It is almost cleanliness before godliness in our business, so this is really quite an advantage. It is the type of system that is used initially to produce the raw silicon that we start with.

In addition to cleanliness however, the mechanism by which the vapor deposition takes place is very important in obtaining the dimensions we are after. I shall try to illustrate this. Please refer to Figure 4 which illustrates a more-or-less idealized structure. If the device illustrated an alloyed structure, we would start on the top surface - alloy in an emitter - and start on the bottom surface,and alloy in a collector. What was left between the alloying on the top and the controlled alloying on the bottom would determine the thickness of the base, and is a very poor control.

In producing a diffused base transistor, we would start on top and produce the base layer downward by diffusion and then start on top again and produce the emitter layer by diffusion. It is suddenly realized that a very thick collector region is not needed and the unnecessary thickness is costing us performance in terms of both saturation voltage, being used as a switch, and its high-frequency capability. So we decide that we will make the collector contact by diffusing in from the bottom. Ironically, we are back to the same problem we had in alloy transistors where we again have a relatively poor control in the thickness in the structure of the entire wafer.

In the case of epitaxial growth we can start with the collector contact, and it can be any thickness desired, and we can deposit on it precisely the amount that we want of collector region. If we were doing the process today, we would build the maximum thickness and then go back to our old-fashioned process and diffuse in the base layer and the emitter because we know how to do that. The first feature that we would wish to exploit from epitaxy is the ability to make the controlled collector region. As time goes on though, I am certain that we will also learn the very desirable situation that

each one of these critical regions, which are collector, base, and emitter, is deposited independent of the others, and therefore its thickness and its resistivity is not a function of those of the other layers. This is essentially an ideal process from this point of view. Furthermore, when we get smart enough, we presumably can vary the doping as we go, and get any profile that we want. In the alloy and the diffusion processes we cannot do just anything we want to do. There are only certain distributions which Mother Nature will give us, and we have to live with them. They are pretty good, but they are far from ideal in many cases.

The way the doping is accomplished in the epitaxial process is to mix a compound of a doping material such as boron trichloride (if you want a P-type region) in with the silicon gas, such as silicon tetrachloride, and cause the silicon and boron to deposit simultaneously. I might point out that an optical diode structure can be made by the epitaxial growth process. One desires in a diode a thin base region with heavily doped contacts on both sides again to reduce the resistance.

Ironically and interestingly enough a type of transistor (intrinsic barrier) which was invented over five years ago by Jim Early at Bell Labs can now really be made effective where it could not really be by the processes previously available. This is just one example of the kinds of devices of which people have been thinking for years but which they simply have not been able to produce because they did not have complete flexibility of both thickness and doping dimensions.

After we make the semiconductor structure there remains the routine job of making contacts which are unfortunately often taken for granted. One might think that they would be easy, but unfortunately, making contacts has proven to be a very difficult job in our field. I suppose that this is because we keep trying to do more than we really can carry out.

By contact, I mean the metal strips (see Figure 4) that are deposited on the various regions and which must be very close to the junction to cut down resistance. One of the problems is that they must be ohmic to a very high degree - not rectifying and not introducing unnecessary resistance. One reason the problem is difficult is that sometimes the semiconductor we are trying to contact is 1, 2, or perhaps 3 mils thick. Three mils is about the diameter of a human hair. Often the junction is 1 or perhaps 2 wave lengths of light in thickness.

Contacts are made by a number of processes often by evaporation and typically with a small amount of alloying to insure the ohmic nature of the contact; chemical plating is also widely used. We are making progress but contacts are a problem that is by no means completely solved. Once one gets the contact, the job is easier. It should be soldered or one can use a thermocompressive bond (again discovered at Bell Labs). In a thermocompressive bond the element is heated slightly together with a wiring contact and a knife edge applied at high pressure to make the bond.

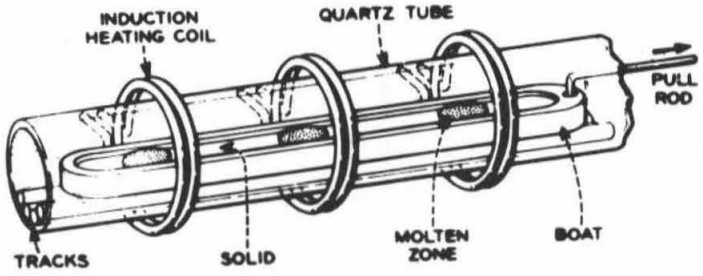
Figure 10 is an example of connections made by thermocompression bonding. It is difficult to make an aesthetic and viewable picture of one of these devices. (Figure 10 shows a typical fast computer transistor). The wires going to the base and the emitter are about half of the diameter of a human hair.

Figure 11 shows transistors with the linear structure. The metal contacts can be clearly seen. The emitter is interleaved back and forth and is surrounded by base contacts. It is a cone-type structure, and the base contact fingers go between the emitter contact fingers to reduce the parasitic resistance. Some of these transistors are experimental types and really involve biting off more than we can chew at present. However, it does show the flexibility of the processes and what one gets into when one tries to make efficient lead contacts.

Finally after one has taken all of the steps I have described, the device should be put in a condition in which the customer can use it and it will stay in good condition. The package must provide for good heat transfer. For example, if one makes a high-frequency device, it should be small, and the only real limitation in size reduction is the ability to get out heat. Good electrical connection to it should be maintained to insure long life. One must support the device so nothing happens to it, and most important, keep out the contaminants. The importance of the surface has been illustrated already by the number of references to it. The surface of the semiconductor device is truly its Achilles heel, and a great deal of effort has gone into better understanding and better control of surface properties, and I think a considerable amount of progress is being made.

When one recognizes that perhaps one out of a thousand atoms on the surface (if it is the wrong kind) will ruin the junction, this is not very many. I mean a thousandth of a monolayer of atoms is enough to ruin the device. This again puts us into an area of purity and control that is quite new. We have not as yet completely learned to deal with it. Certainly, scrupulous cleanliness reduces the problem and there is an effort going on now in a number of companies to improve the ability to maintain surface properties by applying bonded layers to the surfaces. An example is growing an oxide directly from the silicon by reacting the device with an oxidizing atmosphere. This can be done before or after the junction is formed, but the idea is to get an oxide layer over it which has the effect of masking the bare silicon surface.

In summary, semiconductor technology is developed around two requirements, and what I have described has been our attempts to answer these requirements. The first is the almost incredibly small dimensions. Thickness dimensions may be one wave length of light or thereabouts. Lateral dimensions may be of the order of 1 to 2 mils. The other requirement is for material perfection - a purity and cleanliness which would be ridiculous if they were not so essential.



Axially aligned induction heaters are shown in a diagram of the zone melting equipment.

Figure 1

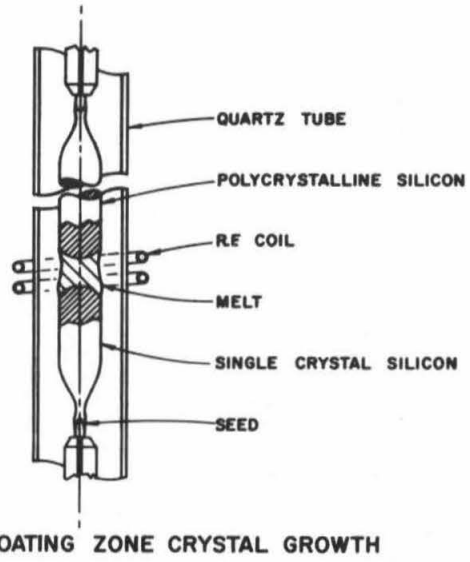


Figure 2

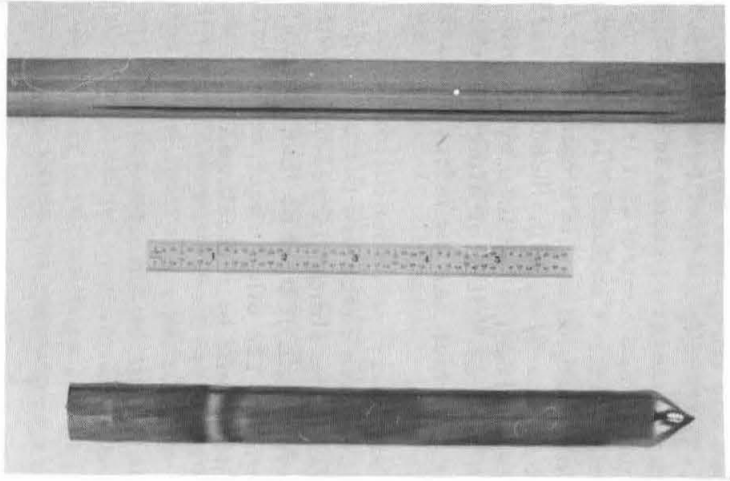
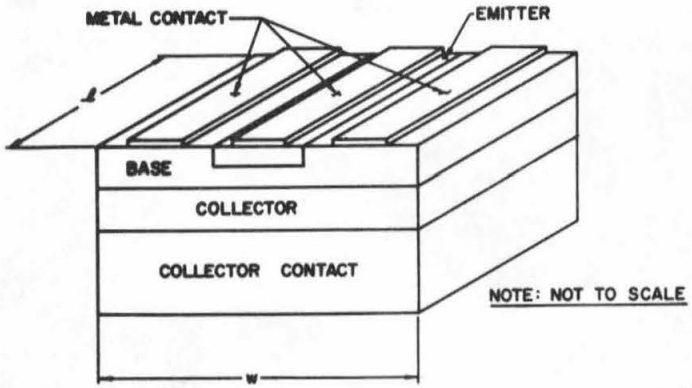


Figure 3



ISOMETRIC DRAWING OF A TYPICAL
HIGH FREQUENCY TRANSISTOR

Figure 4

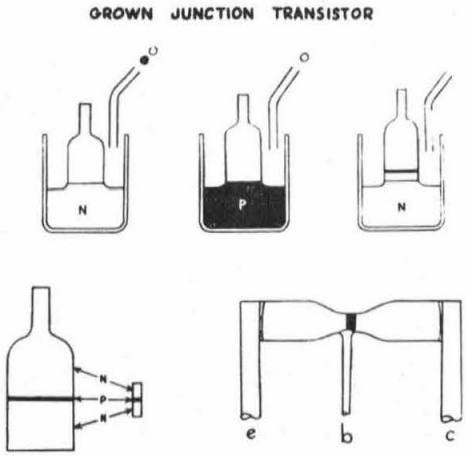


Figure 5

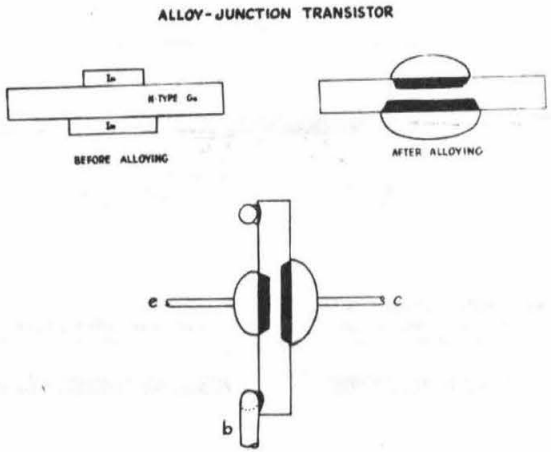


Figure 6

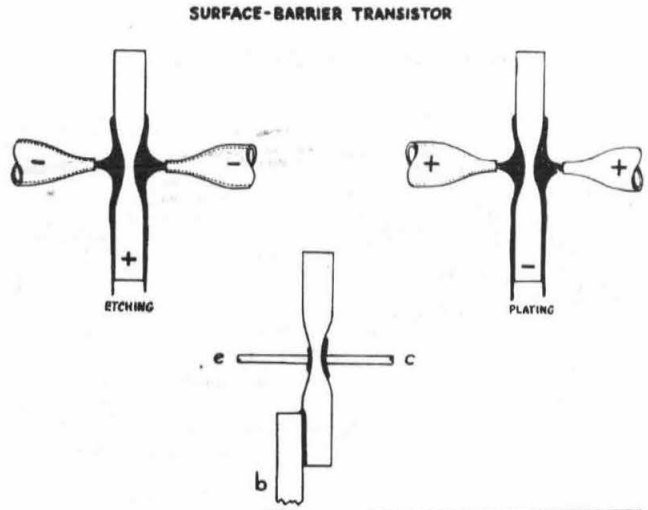


Figure 7

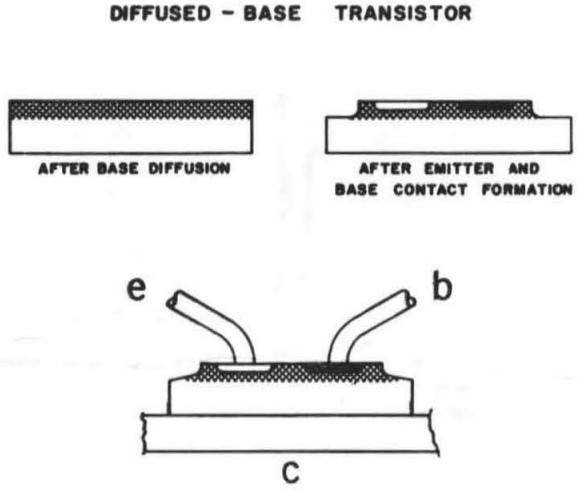


Figure 8

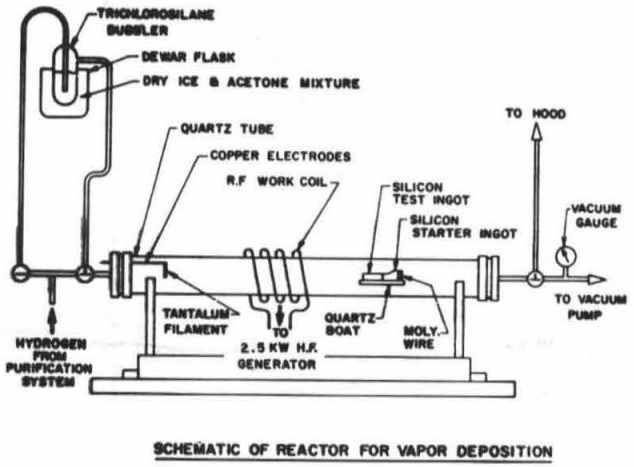


Figure 9

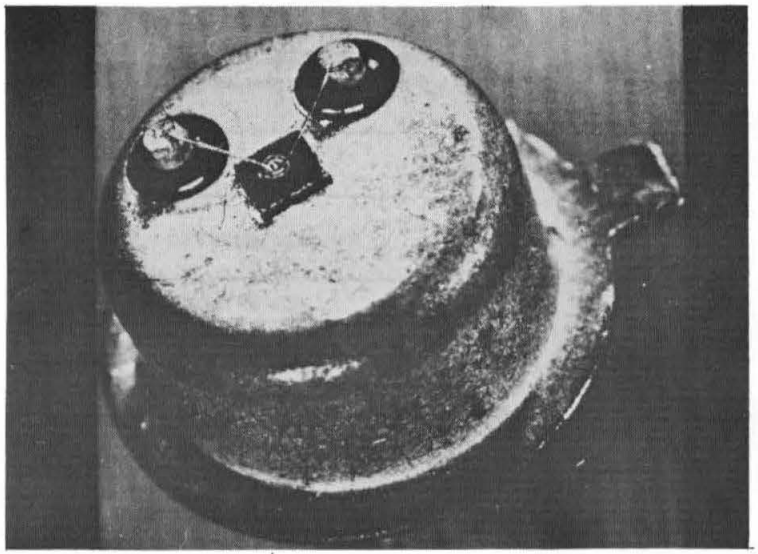


Figure 10

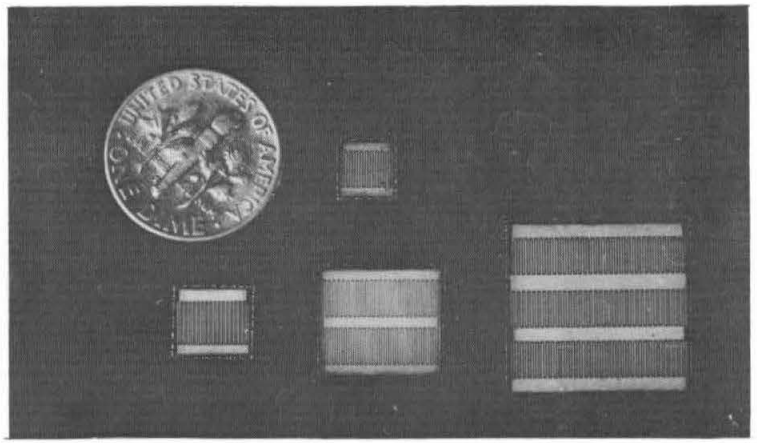


Figure 11

THE GROWTH AND POTENTIAL APPLICATION OF DENDRITIC CRYSTALS

John K. Hulm
Westinghouse Electric Corporation

As you have seen from the previous talks in this seminar, a great deal of solid state technology, as we know it at the present time, depends heavily on the synthesis of single crystals. From the properties of the single crystal devices which we now employ, and also of those which are under development, it seems likely that the single crystals required in the future will include not only very highly perfect homogeneous crystals, but also lattice systems in which controlled variations of physical and chemical purities are deliberately introduced. To satisfy such requirements, we must not only improve the existing growth techniques, we must also search for completely new methods of synthesizing crystals.

In the semiconductor field, much of the existing manufacturing technology utilizes crystals grown from the melt. Recently we have seen the introduction of vapor growth techniques. As the previous speaker has pointed out, the vapor growth technique provides us with a new flexibility and power to introduce exotic variations of composition to a semiconductor crystal.

Should we regard the development of melt growing to have reached its ultimate potential?

I would like to try to convince you that this is not the case. Indeed, innovation is still possible in this area. My thesis involves the examination of so-called dendritic growth, a phenomenon that was regarded as a nuisance in the semiconductor field, at least in much of the early melt growth work. The occurrence of dendrites due to supercooling in the melt frequently prevented one from getting a nice, big single crystal. If, however, one takes the opposite approach of trying to encourage rather than to suppress the dendrites, nature rewards us with some very interesting phenomena.

When a liquid is cooled carefully to its normal freezing temperature, in the absence of suitable nuclei for crystal growth, the liquid state can be preserved at temperatures well below the freezing point. This phenomenon of supercooling can extend as much as 100° (centrigrade degrees) or more below the freezing point. When freezing eventually does occur, it is likely to be extremely rapid since the supercooled liquid provides an excellent sink for the absorption of latent heat of solidification.

When supercooled liquids freeze, they frequently do so by the formation of elaborate crystal forms. The word dendrite is usually used here. It is derived from the Greek word dendros, which means tree. Dendritic growth seems to occur with the largest possible ratio of surface to volume for the crystals which are formed. This is to allow the latent heat of solidification to be absorbed by the liquid, with the minimum possible reduction of the supercooling driving force. Thus, dendrites usually grow as needle-like or plate-like crystals, with many different branches. Everyone has seen the dendritic crystals of ice which form when a supercooled puddle of water freezes.

Dendritic growth in semiconductors appears to have been first investigated seriously by Billig in England in the early 1950's. From a supercooled germanium liquid at 20 centrigrade degrees below the freezing point, Billig grew elaborately branched germanium crystals of the type shown in Figure 1. The branched germanium crystal

has a tree-like structure with many branches as indicated on the bottom of the figure. Billig found that if he took one of these dendrite branches and used it as a seed, he could cause the branch to grow singly, without very much subsidiary branching, as shown on the top of the figure. One might call this controlled dendritic growth.

It seemed to us that this type of growth offered certain attractive possibilities for device improvement, and we have done quite a bit of research work on it at Westinghouse in the past few years.

I would like to discuss the salient features of the growth process as far as we understand it at the present time.

Figure 2 shows a ribbon of germanium dendrite being pulled by a continuous process in our laboratory. One can see the molten germanium surface, the dendritic ribbon, the meniscus of the liquid, and the reflection of the ribbon on the liquid surface.

Figure 3 shows a comparison of this type of germanium ribbon with a typical ingot of germanium pulled by the well-known Czochralski technique. I think it would be illuminating to briefly compare and contrast the two processes.

If there is a molten pool of semiconductor material contained in a crucible with a solid rod of the same material extended into the pool, then this would illustrate the typical Czochralski geometry. Power is generated usually in some suitable kind of crucible graphite containing the liquid, and this power is flowing into the system. This power is also flowing out by radiation from the surface of the molten liquid, by conduction through the solid, and by radiation from the surface of the solid. If the solid ingot is being drawn from the melt, in a crystal growing process, it is necessary to consider the possible ways of removing the additional energy introduced into the system by the liberated latent heat of fusion. This latent heat can be removed by conduction up the rod. A fairly good approximation to the limiting conditions is the limiting velocity of pulling of the crystal which was obtained by Billig himself. He calculated this on the assumption that the surface of the melt is maintained at the melting temperature and that the thermal gradient in the solid is determined entirely by the flow of the latent heat. From this consideration along with the inclusion of certain assumptions about radiation from the surface and so on, he concluded that the maximum pulling speed would be roughly $\frac{.01}{\sqrt{r}}$ cm/sec., where r is the radius of the ingot. If one substitutes the typical size of Czochralski ingot, which would be, say, a centimeter radius, then the pulling velocity is .01 cm/sec, which turns out to be about 15 inches per hour.

Actual pulling speeds of these ingots are usually about an order of magnitude lower than this, and this can probably be accounted for by the fact that in this calculation Billig assumed black body radiation, and one does not really have this condition; the emissivity is lower than unity, and also there still is some power from the melt which is flowing into the ingot, and so all the latent heat cannot be removed. This makes a contribution to the film gradient. Most of these factors would lower the pulling speed.

For a very thin ingot, a flat ribbon, such as one obtains in dendritic growth, it is necessary to consider the thickness of the ribbon, the thinnest dimension, which is of the order of 10 mils. This thickness is the effective radius "r" of the ribbon. Thus, for a dendrite one calculates a pulling velocity of .05 cm/sec. The pulling velocity was calculated by assuming that the dendrite is pulled while the melt, which is not considered to be supercooled, is maintained at a constant temperature. The actual pulling velocities of dendrite are much greater than this.

Figure 4 shows a plot of the speed of pulling in cm/sec. This is the range of supercooling for a typical germanium dendrite. Actually, we usually work at higher pulling velocities. The plot is for low supercooling with the melt at 10 or 20 centrigrade degrees below the freezing point. It is possible to obtain pulling velocities of a centimeter a second or more. It is clear that the assumptions here cannot explain the high speed of dendritic growth. It is not possible to account for the removal of latent heat by just assuming conduction and radiation.

Probably what happens is that most of the latent heat is absorbed in the supercooled melt. This is not obvious either, because if one had a source here which was generating heat in a stationary condition, one would not be able, even with a supercooling of 10 centigrade degrees, to remove a very high amount of heat in this direction. The reason for this is that there is not enough driving force in thermal gradient. However, one could think about the dendrites as needles growing very quickly through a liquid. Imagine some needle-shaped crystal, moving at high speed through a liquid semiconductor, and penetrating new regions of supercooling, so that, in effect the liquid is flowing rapidly past the dendrite. A disturbed condition would exist in which new areas of supercooling are made available. The removal of heat can be explained on this basis.

It turns out that very probably the limiting speed of growth is determined by the radius of curvature of the needle-shaped crystal. The sharper the needle, the faster it can grow.

This process of growth requires a mechanism for the deposition of atoms at the tip of the needle. To visualize this, it is necessary to investigate the crystallography of the dendrite.

Figure 5 shows in a very schematic manner the crystallographic form of the dendritic growth in the diamond lattice that we have examined most in our laboratories. Suppose it is assumed that the dendritic ribbon extends below the melt surface in the form of a wedge rather than a needle. Instead of letting the wedge grow downwards, through the supercooled liquid, as shown here, one deliberately pulls the ribbon out upwards at such a rate that the tip of the growing dendrite is stationary in space. But, by doing this, material is removed so that new material has to flow in continously to replace it. The results of this operation are illustrated in the figure.

The point surfaces of the ribbon are (111) planes, which are the closest planes in the diamond lattice. We have found that the crystal usually contains - always contains - one or more central twin planes parallel to the (111) faces, and these twin planes, we think, play a vital role in the crystal growth. Growth occurs in a (211) direction. The details of the actual growth of the dendrite are more complicated than is shown in Figure 5. This can be examined by jerking the dendrite out of the melt suddenly, at very high speed. Very sudden jerks will break the dendrites. It is necessary to reach a compromising situation where it is possible to jerk the crystal out without leaving too much liquid on the end. From the jerked dendrite, it is possible to obtain some indication of the actual growth surfaces.

Figure 6 is an enlarged photograph of such a jerked dendrite. I will describe the growth process in three distinct parts. First, the extension of the tip region, which is a very small region on this slide; the second part is the thickening of the dendrite; and thirdly, the lateral growth of the dendrite.

Attention is focussed for a moment on the extreme tip of the dendrite, where the twin planes protrude, into the melt. It is very difficult to get a good photograph of this tip that is growing since it seems to pick up more liquid than the rest of the system. We believe that initial nucleation of new growth layers occurs at the tip, due to the existence

of re-entrant corners between other (111) planes, which intersect at the central twin plane. This is shown schematically in Figure 7.

The dendrite is growing in the direction shown by the arrow and these are other (111) planes which intersect at the central twin planes in the crystal, that is if there is a single twin plane. If there is a single twin plane, (Figure 7(a) which tends to occur, it can be shown quite readily that the corners grow out and rapid growth would stop. In this respect the external corners grow out while the nucleation sites are being lost.

However, if there are two twin planes, such as illustrated in Figure 7(c), the corners may try to grow out in this one system, but the growth layers generate new corners, new sites of nucleation, like II, which cause the growth to continue on this plane. The situation is probably much more complex in practice. There may be a whole series of growth planes in motion at once, such as shown in part (d) of Figure 7. The type of motion is somewhat familiar to the screw dislocation motion around the direction of growth, in that it is a spiral growth process.

It has been confirmed by several people; for example, Wagner at Bell Labs, that two or more closely spaced twin planes are essential for the continuous growth propagation of these dendrites. These twin planes, which form a central core of the dendrite, usually are a few microns thick, and less than 10% of ribbon thickness. The second part of the growth process is the thickening of the dendrite. This occurs immediately behind the tip. A magnified version of this thickening region is shown in Figure 8. There is a terrace structure of growth steps usually bounded by many (111) or close-packed planes. For an odd number of twin planes in the core, the terrace structure is likely to be identical on opposite faces of the dendrite. For an even number of twin planes, from the morphology, it would be expected that the opposite faces would be different. This is shown in Figure 9. The twin planes in this case are marked by arrows. Those are the original twin planes of the growing tip, and this point is widened quite considerably at the section as shown by the line where this cross-section was taken. The other mounts on this cross-section are due to etching and they represent some segregation of impurities.

In the thickening region, it is quite common for the growth terraces to exhibit a deep groove parallel to the direction of growth. This is shown on the right-hand side bisecting the terraces. Alternatively, the single growth steps themselves can be quite heavily grooved, as shown in Figure 10. This grooving is not understood completely, but I mention it because it has a considerable influence upon the perfection of the final dendrite, which will be discussed later.

When the dendrite has attained its final thickness, which is controlled by quite a number of factors, one of which is probably the radius of curvature of the growing tip, it enters the third region of growth. In this third region lateral expansion occurs. The lateral expansion is a very interesting region. It is shown in a very schematic fashion in Figure 11.

When the thickening of the dendrite is complete, lateral growth begins in the (011) direction, but the actual planes that form the lateral growth region, the growing planes, are (111) faces once again. These develop, we believe, in the thickening region. When complete thickness is attained, the crystal starts to grow out sideways. In this kind of mechanism the arms shoot out in a sideways fashion to form a structure. There is a hollow space between the arms, as is shown in Figure 11. These hollow spaces, of course, contain liquid when the crystal is below the melt surface, and the liquid in the spaces eventually solidifies, as the ribbon is removed from the melt to obtain a solid ribbon.

Frequently the four lateral growth arms which form what we have begun to call the H-structure suppress the growth of the central core of the dendrite, and sometimes the twin planes in the central core grow outwards. They have been drawn growing out a short distance. Sometimes they actually grow out right to the edge of the dendrite during this process. Sometimes they are suppressed completely and simply stay in the center.

Crystallization in the spaces between these H-arms seems to occur in a variety of different ways, depending on thermal conditions on the doping level of the material. Examples of extremes of behavior are shown in Figures 12A and 12B. Figure 12A shows a section of the H-arm structure taken on a crystal which was jerked out of a melt and the liquid was left behind, which is not usual. Sometimes surface tension holds it in place.

Figure 12B shows the two extremes of behavior of solidification in the H-arms. Again there is quite complicated segregation in these materials. However, in Figure 12B (b) the twin planes went right out to the edge of the material, and in slide 12B (a) they were suppressed. It is quite possible, if you have an even number of twin planes, for the growth to be pinched off. If you have an odd number, you may get just a random boundary which wanders out to the edge of the crystal.

The sawtooth pattern which one observes along the edge of the H-arm structure is preserved in the final ribbon. Usually, the dendritic ribbon of germanium is about ten or twenty times the thickness of the H-arms and the thicknesses range in our experiments from something like 1/1000 of an inch up to 20/1000 of an inch. The white faces of the germanium ribbon appear to the eye to be very shiny and absolutely mirror-smooth, except for the presence of a series of more-or-less pronounced curved lines which are shown in Figure 13. These lines are always concave downwards towards the melt. In other words, the pulling direction is this way and the melt is downwards. We know that they consist of a series of rather shallow steps, which range in height from a few Angstroms up to 5000A. It is believed that the steps are formed during the final stage of growth. Even though lateral growth is dominant in forming the H-arm structure, growth layers do occasionally nucleate on the flat (111) faces of the ribbon. Such layers apparently spread out to the melt surface, and they cannot grow any further because they run out of liquid. The melt surface is curved by surface tension. One observes that as the dendrite is pulled upwards, the meniscus of the liquid also tends to move upwards on the flat face for perhaps a millimeter or so, and then drops back, repeating this process irregularly.

Figure 14 shows in a schematic way what might be happening in this. The growth layer is nucleated on the surface of the dendrite below the melt and goes up to the melt surface and forms a kind of a corner at this point. It cannot grow any further because it has no more liquid. The energy situation is changed so that you can pull the meniscus up above the melt, but at some point it will break away and drop back to its normal position and leave this growth step. This stick-slip mechanism seems like a reasonable description for the growth steps on the surface.

Dendrites in which growth is continued for a few inches will usually exhibit all the structures which I have so far described. However, since growth lengths of thousands of inches per hour are readily obtained, it is possible to grow the material in greater lengths. I believe that the maximum length that we have grown up to the present time are several hundreds of feet. This is perhaps the longest single crystal that has ever been grown. To handle this type of material, we utilize the fact that the dendrites are quite flexible and can be curved to a radius of one foot or less.

In our continuous pullers, dendrites coiled on a dismountable disc can be handled in much the same fashion as wire. Figure 15 shows the type of system we use for

experimental purposes. One can simply establish an exit lock from a furnace with a continuous stream of inert gas flowing out to prevent oxidation of the melt and pull the dendrite directly out. However, for many considerations, such as prevention of surface contamination, and maintainage of controlled vapor pressures, it is necessary to do the coiling in the furnace atmosphere. Figure 15 shows one type of puller, which is called a low-pressure puller. It works only up to atmospheric pressure. The dendrite is grown in the induction furnace and is pulled up through a lock, and then is coiled on the drum which runs on the inside of this furnace. It has a lucite plate on it so that one can have a very good view of what is happening during the coiling. We usually wrap a tape in with the dendrite to prevent the dendrite from uncoiling if any breakage should occur.

Figure 16 shows a puller for higher pressures which works on very much the same principle, except it will stand up to about 20 atmospheres.

In pulling continuous dendrites for semiconductor applications, it is desirable to obtain proper control of the physical dimensions. It is also desirable to maximize the crystal perfection and to control the distribution of impurities. A few comments on our experience in these areas seem appropriate.

One of the major problems of dimension control stems from the tendency shown by most dendrites to get steadily thicker, due to the cumulative effect of the curve growth steps on the ribbon face. It seems unlikely that such growth steps can be completely eliminated. However, we know that it is possible to greatly reduce their chances of formation. We have produced germanium dendrites of a few mills thickness, with a variation of thickness of less than 1/10 of a mill over a 20 foot length. Some typical curves of thickness versus length are shown in Figure 17.

With respect to dislocations, it is possible under adverse growing conditions to have the dendrite faces virtually covered with dislocations. On the other hand, under favorable conditions it is possible to greatly reduce the dislocation count. The areas in which dislocations typically occur are shown schematically in Figure 18.

If the spine in the thickening region (I) extends into the lateral growth region, there is a danger of liquid entrapment in a hole in the surface, and you seem to get freezing over the surface first. The liquid below, when it finally freezes, expands and produces a very bad deformation of the material. We have been able to eliminate this.

The second region of dislocations is out near the end of the lateral growth region, and this is a pretty hard thing to eliminate, but the density of dislocations is usually much smaller here.

There is a third possible region close to the center where there is a heavy dislocation density if there had been an entrapment of liquid. This is very similar to the mechanism in the second region. However, in this case liquid is trapped between the H-arms, and somehow if this liquid is pinched off by growth inwards, then the same kind of deformation occurs and a large concentration of dislocations is formed. We have been able to eliminate this by modifying the growth conditions.

One might expect a direct connection between minority carrier lifetime and the occurrence of dislocations. This is borne out by experience with the early dendritic material, which was heavily dislocated. We observed lifetimes ranging from one microsecond up to 15 microseconds. As we are able to reduce the dislocation count, the lifetime is steadily improved. In present dendrites between 5 and 10 mills thickness, we observe lifetimes around 40 microseconds. It seems probable that this value is mainly determined by recombination at the external surfaces of the dendrite, and that the actual bulk lifetime of the material is considerably in excess of 40 microseconds.

It is of interest to note that from a separate series of experiments on twin boundaries, it is possible for us to conclude that the recombination velocity associated with the twin planes is very small. We feel that the presence of the twin planes is not disadvantageous, even in the case of multiple twinning as far as the utilization of the material for devices.

The fairly complicated mode of crystal growth which I have described in the H-structure of the dendrite leads one to expect a non-uniform distribution of impurities in the material. However, under carefully controlled conditions of growth, we have no difficulty in achieving constant values of the cross-sectionally average resistivity over quite long lengths of the material. For example, in several hundred feet of germanium, we have been able to control the average, the cross-sectional resistivity, to about 10 per cent over the entire length. This is true for various doping levels. Typical curves of the cross-sectionally averaged resistivity versus length are shown in Figure 19. The resistivity is quite uniform over the entire length.

As far as local non-uniformities are concerned, one would expect from the nature of the growth a greater resistivity across the thickness of the ribbon - for the peak of the center, approximately in the same position as the slots between the H-arms. For an impurity with a segregation coefficient less than unity, one might attribute this to a piling up of impurity in the slot, which is the last part of the dendrite to solidify. However, I have said that solidification in the slot appears to occur in a variety of different ways, depending on thermal geometry and other factors.

When freezing takes place from the cross-structure outwards, this seems to produce less piling up of impurities than when it occurs inwards from the H-arms. The ratio of the resistivity of the center line to the resistivity at the outside edge may range from around 2 or 3, to quite large values, depending on the conditions.

Finally, I would like to point out some of the features of dendritic crystals which seem advantageous from the point of view of device clarification and comment on some of our devices experience.

The range of thicknesses in which dendritic ribbons have been grown embraces most of the thicknesses of semiconductor devices used in present production. The use of dendrite would thus have seemed to offer the possibility of eliminating much of the slicing and cutting and orienting involved in the preparation of devices from massive ingots. Moreover, a dendrite's surface does not need to be etched to remove mechanically deformed areas of the type introduced in the conventional slicing and cutting. These advantages could result in a significant reduction in material cost. A more intriguing possibility stems from the hope that by producing a more uniform crystal in a dendritic process than in a large ingot process, one might improve the yield of acceptable devices in manufacturing. With this in mind, we fabricated a substantial number of conventional devices on both germanium and silicon dendrites. These include alloyed and diffused diodes, tunnel diodes, alloyed transistors and junction transistors and diffused based mesa transistors. To illustrate the kind of results that were obtained, I present some yield data for some alloyed junction transistors. These were made from n-type starting material, with a resistivity close to 1/2 ohm centimeter. Both junctions were made from gallium indium alloy, using a mold fusion process. The data refer to a few hundred units in each case. In our particular process, which cannot be regarded as well-optimized, the distribution of current gain was about the same for dendrite and ingot material. As a matter of fact, the yield of acceptable units was slightly lower in the dendritic case.

The acceptable units in this case had a gain greater than 22 at 455 kc/sec. Gain at 455 kc/sec is plotted on Figure 20, against per cent of units. There is a slightly smaller yield in the dendrite case, but the distributions are not too different.

However, as far as the maximum voltage performance for these transistors, the dendritic material proved to be far superior to the ingot, as shown in Figure 21.

Eighty-four per cent of the units passed the collector-emitter-voltage conditions. The design was for 15 volts in this case. These classes shown along the bottom represent conditions on all three voltages, the VC, EVC, and the EV, but the collector-emitter voltages typical in this particular transistor are shown. It is quite remarkable how the dendrite performs in this case. I cannot attribute the improved voltage characteristics of these transistors to any particular property of the dendritic crystal. Perhaps it is a contribution from several factors, such as exact orientation, surface ideology, and so on. However, our experience does suggest that the material is superior in certain respects, and we think that even more improvement might be made in this.

I do not want to close without mentioning another interesting feature of the dendrite which involves the H-structure. If we dope the dendrites with approximately equal amounts of donor and acceptor atoms, at substantially different segregation coefficients, we can have the higher segregation coefficient impurity deposited in excess in the H-arms and the lower segregation value impurity trapped in the slot. This results in a PN junction configuration. There is such a slot shown in Figure 22.

This is an interesting way of producing PN junctions, and we have been able to produce some quite respectable junctions in this fashion by removing various regions and adding extra contacts with fabricated transistors, four-layer switches and so on. It seems probable that other configurations are possible. For example, if one allows a twin structure to grow out to the edge instead of producing a three-layer object like this, you could produce a five-layer.

To conclude then, we feel that in dendritic growth it is possible to obtain essentially single crystal ribbon with optically flat surfaces in the (111) plane. It seems to us that such surfaces will be of growing importance both for junction fabrication and for use in substrate. If we can produce greater widths, it would seem likely that other applications are possible.

I would like to thank my colleagues at Westinghouse in research and also our Youngwood semiconductor plant for making this talk possible.

Figure 1

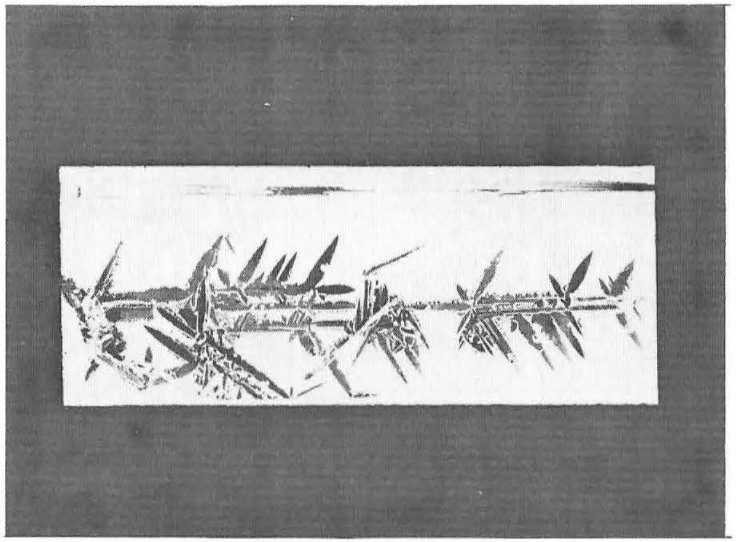
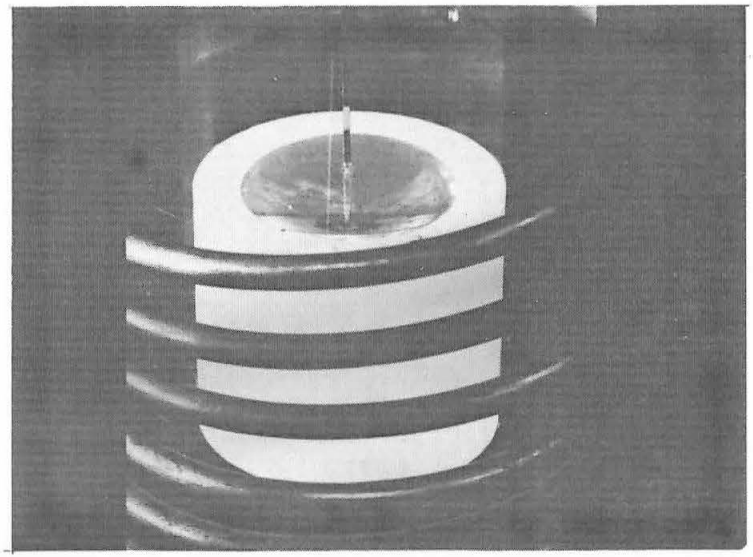


Figure 2



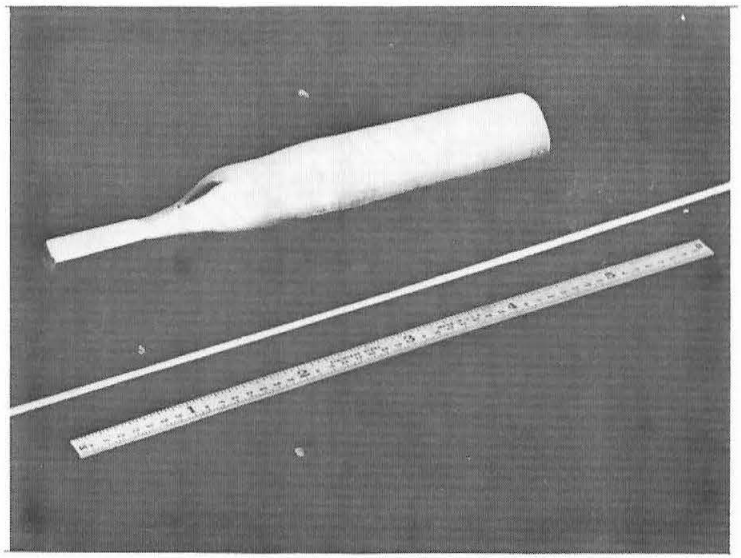
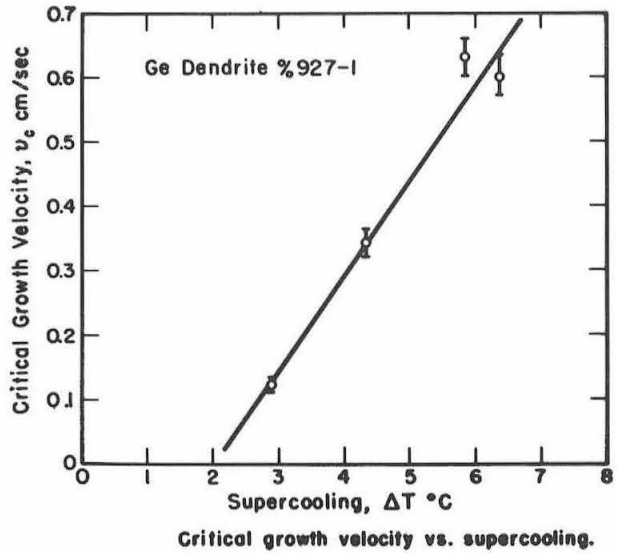


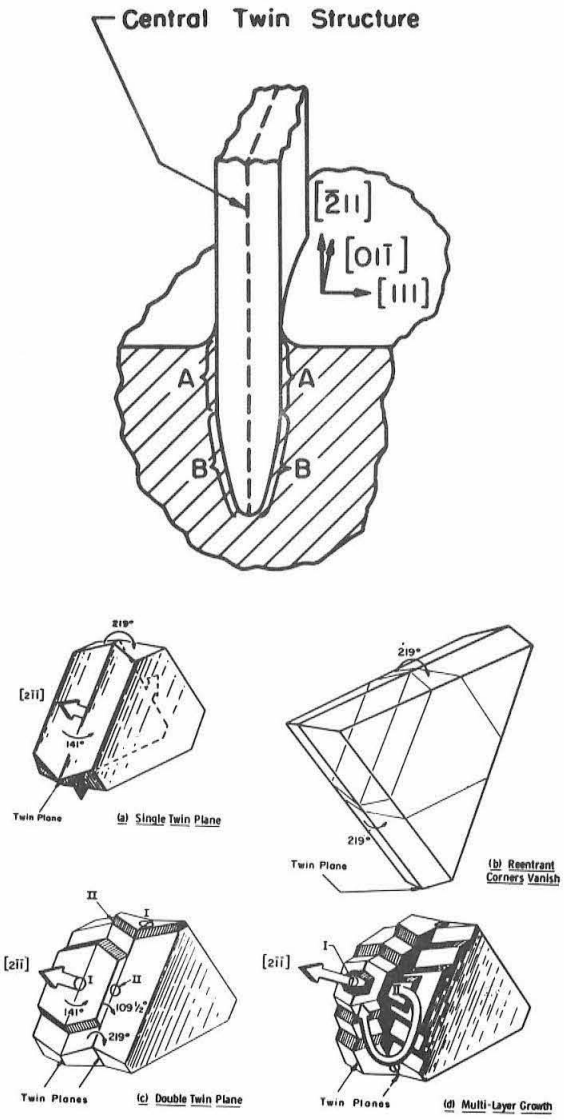
Figure 3



Critical growth velocity vs. supercooling.

Figure 4

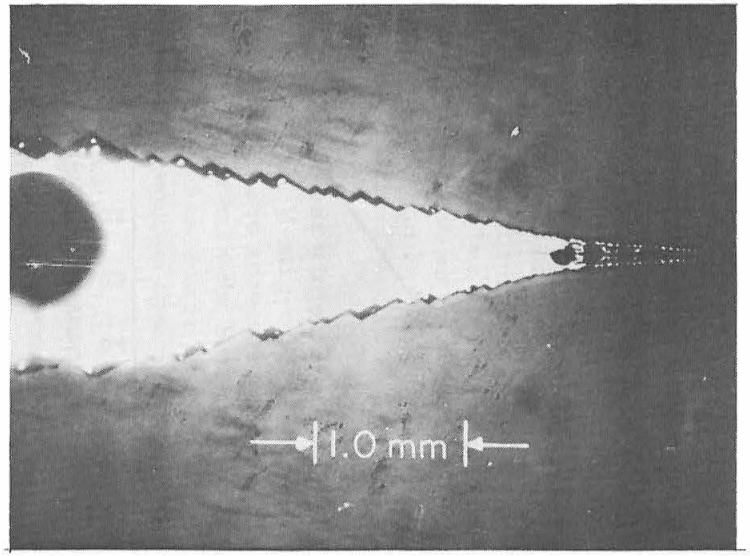
Figure 5

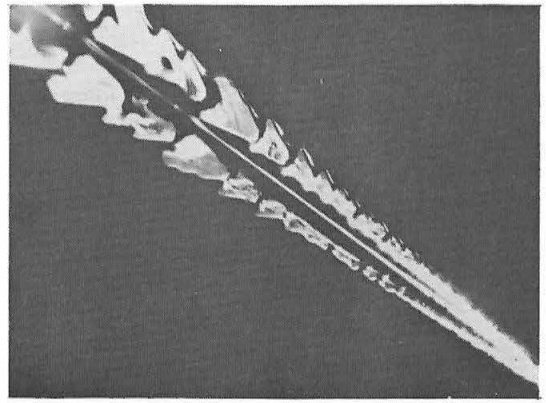


GROWTH MECHANISM AT DENDRITE TIP

Figure 7

Figure 6





Jerked dendrite showing (a) overall growth shape existing under the melt and (b) actual growth steps at a higher magnification.

Figure 8

Figure 9

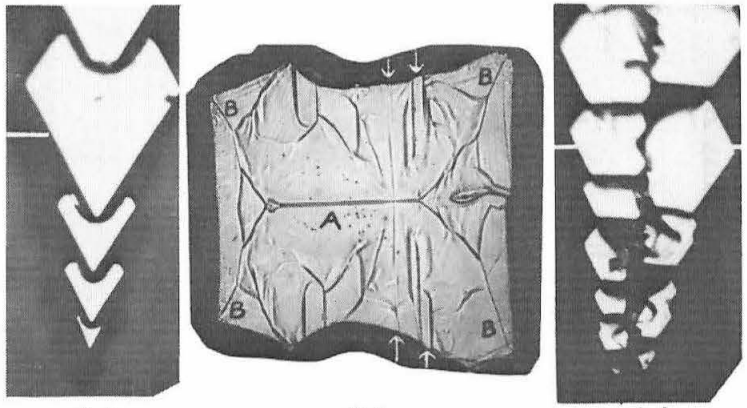
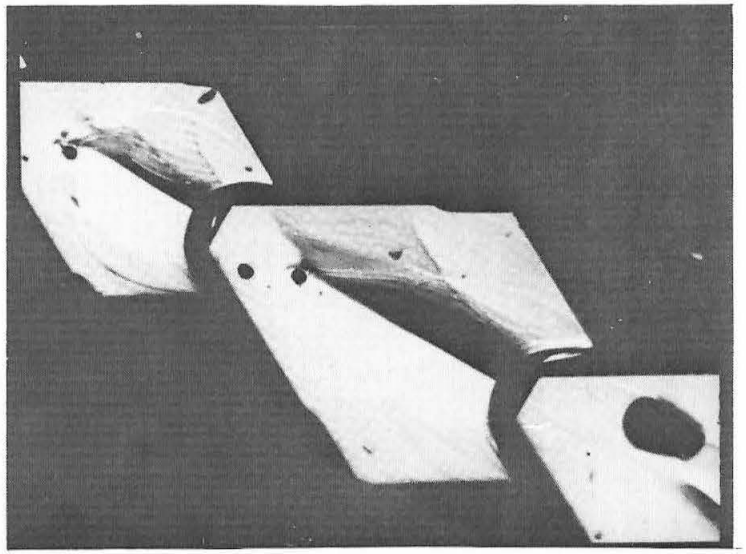
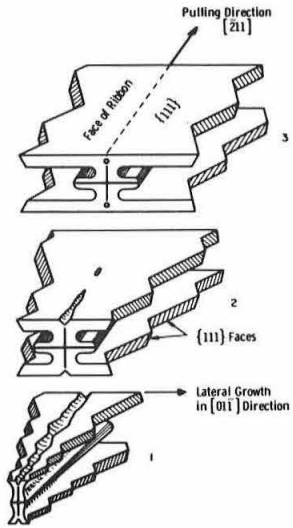


Figure 10





Lateral Growth of Dendrite
Figure 11

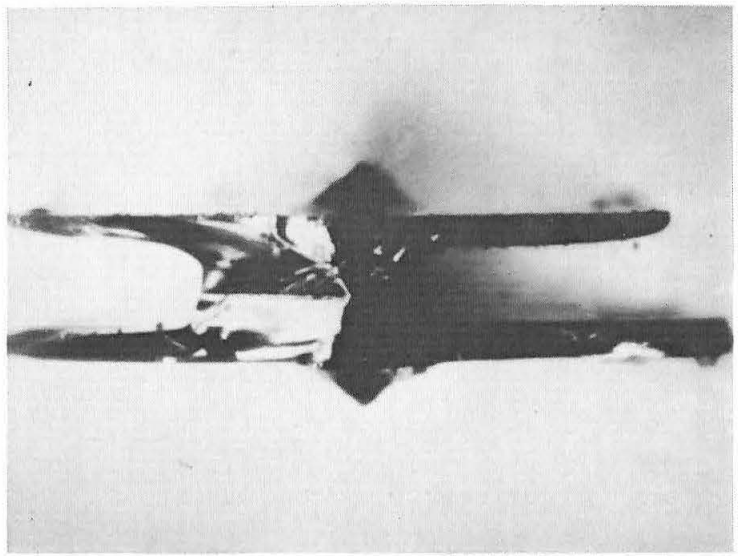


Figure 12 A

Figure 12B

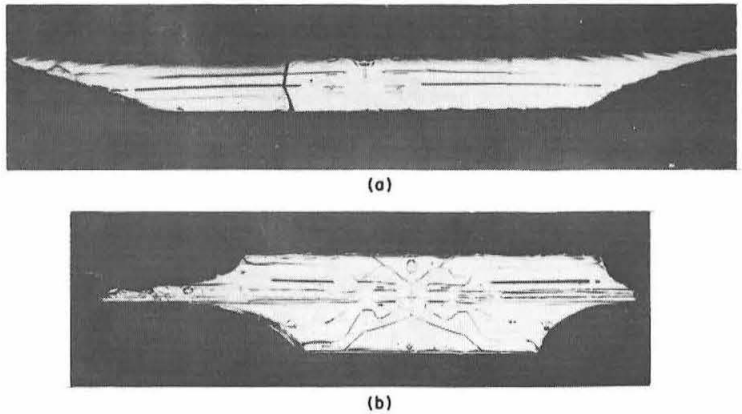
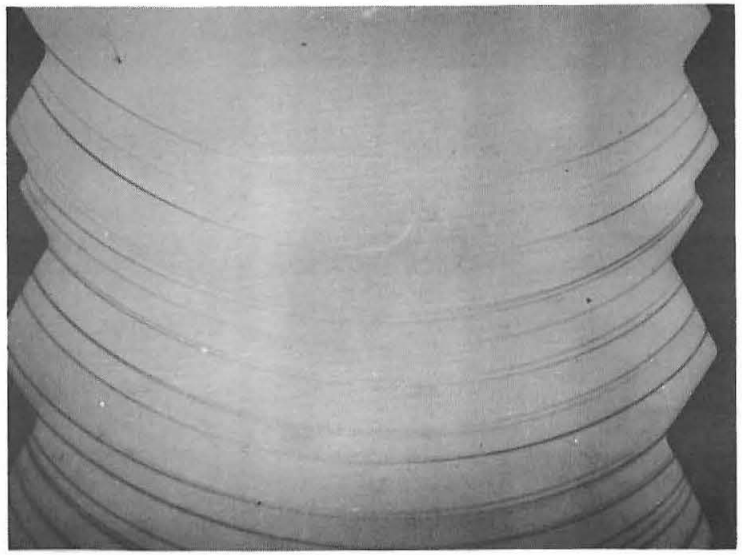
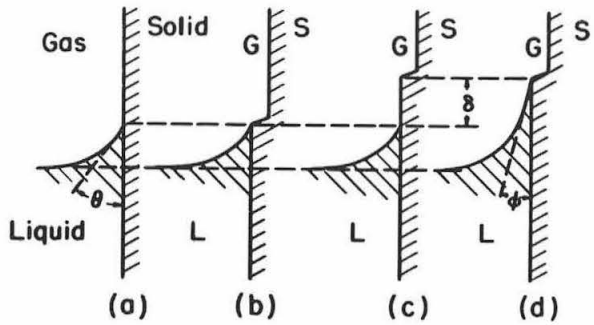


Figure 13





PROBABLE MECHANISM OF STEP FORMATION

Figure 14

Figure 15

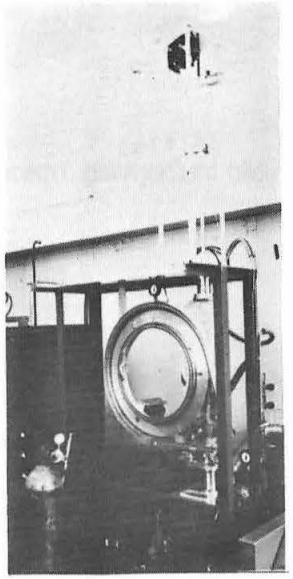
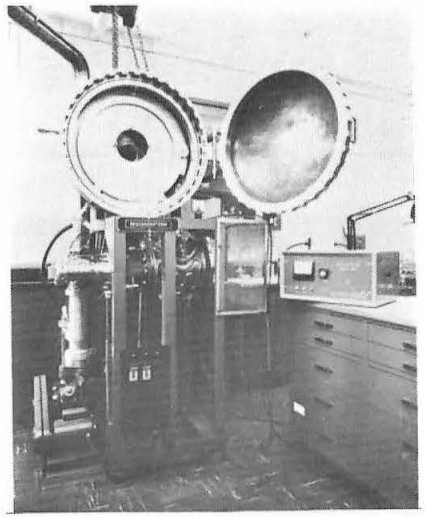
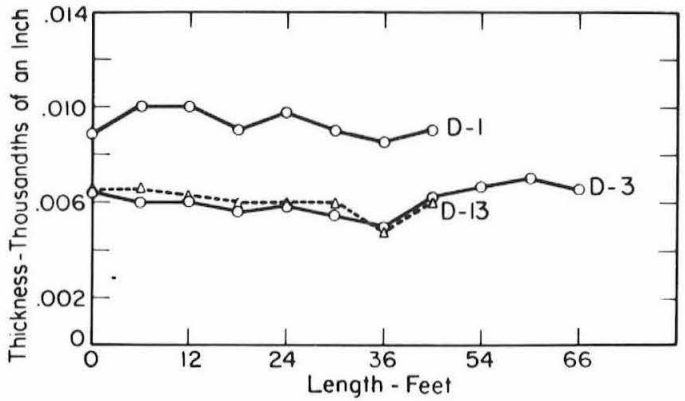


Figure 16

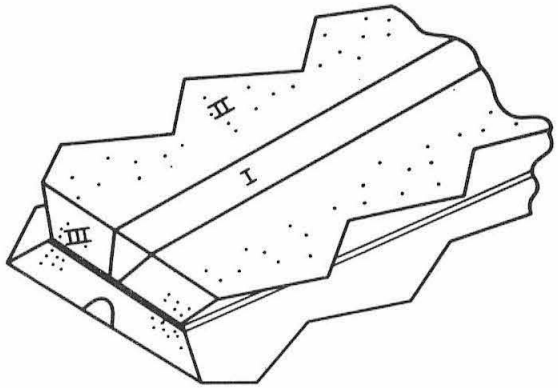




Thickness of continuous lengths of germanium dendrites

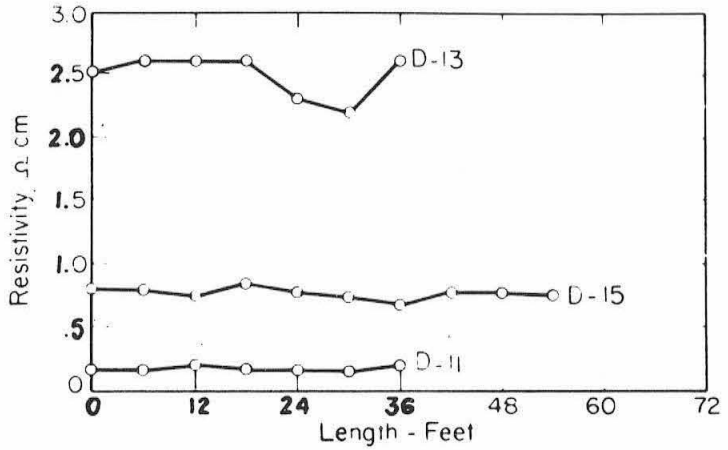
Figure 17

Figure 18



General areas of dislocations.

Figure 19



Resistivity of continuous lengths of germanium dendrites

Gain spread, germanium alloy transistors

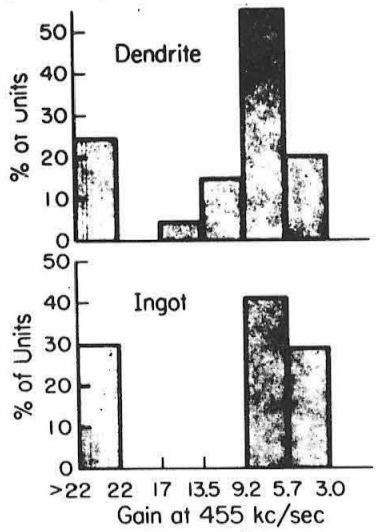


Figure 20

Figure 21

Voltage spread, germanium alloy transistors

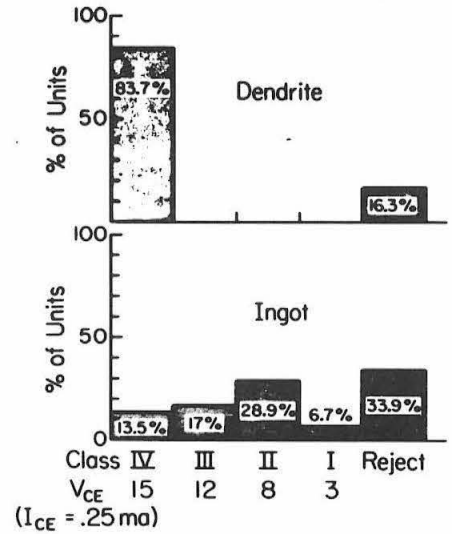
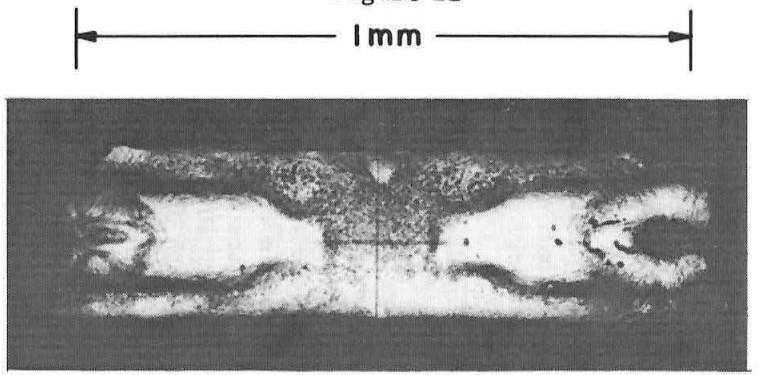


Figure 22



B-Sb doped three zone dendrite

Pulse Electroetched

H-Arm Structure P-Type Material

MICROELECTRONICS

J. Earl Thomas, Jr.
Sylvania Electric Products Inc.

At first glance, there seems to be almost no limit to the possible size reduction of electronic systems. In fact, there is no serious limit to the degree of miniaturization of individual electronic components, but as we shall see, there are some not-so-obvious limits to the possible shrinking of systems.

To keep the discussion specific, most of the considerations will center around electronic systems an order of magnitude larger than the largest systems today, since an order of magnitude increase in the complexity of electronic systems is something real miniaturization might make possible. Since present-day large systems contain as many as 10^5 components, we shall thus concern ourselves with the miniaturization of systems containing on the order of 10^6 components. Also, we shall concern ourselves with electronic systems in which the operations are primarily digital in nature and are performed in ways that are familiar today. That is to say, the electronic functions are to be performed using PN junctions in familiar ways; e.g., transistors, diodes, limiters, together with ordinary resistors, capacitors, and inductors. We shall mention the possibility of performing digital operations in other electronic ways, but we shall have essentially nothing to say about the possibility of miniaturizing circuitry which does not use PN junctions.

The problem that needs answering and to which we shall try to give a preliminary answer is: Which, of all the possible methods for miniaturizing electronics, is the best one to attempt in order to build systems of the type described above? More specifically, can we accomplish all that can really be done in miniaturization using separately mounted and packaged components, or will it be wise to pursue one of the many possible methods whereby component and circuit are fabricated simultaneously or even, as some suggest, should we attempt some highly exotic procedure for fabrication of component, circuit and system simultaneously?

We shall examine briefly the various proposed schemes for microminiaturization. Following that we shall consider limitations placed on miniature systems, particularly on highly complex miniature systems having as many as 10^6 components, by the two dominant considerations of system reliability, on the one hand, and power dissipation with resultant temperature rise in the system, on the other hand.

We shall see that the reliability problem can be overcome to any desired degree by redundant system design provided that the redundancy is cleverly applied. On the other hand, we shall see that problems of power dissipation and temperature rise are not so easily overcome, and indeed, they place an ultimate limit on the degree of miniaturization of electronic systems of the type under consideration here. This limit turns out to be sufficiently large so that it should be possible to approach the limit with separately packaged components, although there may be other reasons than pure miniaturization - reasons such as convenience, cost or reliability - for using multiple packaged components.

Speaking of components, to show the size reduction possible and already achieved, Figure 1 shows a vial containing no less than 1/4 million transistors complete in every respect, including electrical contacts. All that is necessary to get amplification from one of these - at say 100 mc - is to connect three interconnection wires from an external circuit to the three metallic contact regions provided on each transistor. Also, one must provide protection for the unit from atmospheric contamination.

Unfortunately, when these minor problems of connection and protection are taken care of, the picture is changed, as you can see. Figure 2 shows 1000 transistors ready for incorporation into circuits. Each is provided with purely external wires and a protective, hermetically sealed can, and each one occupies at least 10^4 times the minimum volume required by the germanium die.

From Figure 3 we can get a quantitative idea of what could be accomplished if components were the only problem. This is one of the transistors from the 250,000 in the vial, drawn to show the volume of its major parts.

Now to anticipate the discussion somewhat, we shall later see evidence which indicates that an over-all component density of 1000 per cubic inch is the maximum useful density in an electronic system. A look at this transistor will show how amazingly far we have gotten beyond this limit of 10^{-3} cubic inch per component in the case of transistors themselves. Figure 3 is really a die (plural dice) from a mesa transistor. The entire useful electrical function of this die takes place in a small part of the active region shown here at the top of the mesa. Far from occupying 10^{-3} in^3, the entire die is only 10^{-6} in^3, the mesa only 10^{-8} in^3, and the active region itself 10^{-9} in^3. We have not overlooked the volume of the electrical contacts; they are shown and occupy only 10^{-10} in^3.

For comparison the present extreme of the state of the art has been achieved at the Bell Laboratories where J. Early has made a considerable number of units in which the active region is only 4×10^{-11} in^3. Please note, these devices are not made small for smallness sake; rather they are small because it is easier to make them that way, taking into account the electrical job they are required to do. In fact, at the moment, much work is being done to find ways to make them bigger so they will handle more power in order to satisfy the immutable power requirements of many systems for which they are not now suitable.

Thus, as highlighted in Figure 3, the miniaturization of the semiconductor device per se is no problem. The real problem is how to provide electrical interconnection (note that connection and interconnection are not always the same thing), how to provide environmental protection, and above all, how to provide adequate cooling without wasting space. In connection with cooling it is interesting that the tiny device of J. Early's handles - i.e. dissipates - in its fantastically small volume, a power on the order of 100 milliwatts. There appears to be no problem in getting the heat out of the device and transporting it, say, 1/4 inch away, either. The real trouble will come when we have to remove the heat from a million such devices and move the heat 10 feet or more away to get rid of it.

At this point, it is worthwhile to note the odd fact that the miniaturization of active devices has far outstripped the passive. For example, if the active part of a 100 mw transistor can be built into 4×10^{-11} in^3, we may assume that resistors of the same power and size should be possible, but to my knowledge, such a thing has not yet been achieved.

Before going into the heat and reliability problems, let us review the various popular proposed schemes for providing electrical interconnection and environmental protection. These ideas have been discussed many times; and therefore, we will not spend a great deal of time reviewing them. One excellent detailed study has been done by the editorial staff of Electronics.[1]

---

1. Electronics, November 1960.

Figures 4 through 9 show the main features of the commonly proposed methods of miniaturization. In the lower right-hand corner of each figure, a factor is given for the size reduction possible in electronic systems as claimed by the proponents of each miniaturization method. The size reduction figures are drawn largely from a survey done by the U.S. Navy.[2] At the lower left of each figure, a date is given for possible mass production of electronic systems by each of the proposed methods. The dates are drawn primarily from a market survey prepared by the P. R. Mallory Company.[3]

In the survey many people were asked when they expected each of these miniaturization methods to become important in production and the dates given here are essentially the average answer given for each method.

Also, in each of the figures two groups of companies are listed: the first group being those companies that are manufacturing components for this method of construction, and the second group being those organizations active in the assembly of finished systems.

In Figure 4 we see that systems volume can be saved by using flat transistors and other components in close contact with printed circuit boards so that the printed circuit boards can be placed closer together. Alternately, small diameter components can be placed side by side on a printed circuit board. In order to save area on the board such components can be mounted perpendicular to it.

The mini-weld system of assembly, also shown in Figure 4, is particularly worthy of note and is almost certain to be one of the primary forms of miniature electronics for a long time to come. In mini-weld, the components are ideally relatively long and narrow with electrical leads coming out the ends. The glass diode is an excellent component for this method of assembly as presently being made, but the transistor needs a special design. The components are placed side by side much as the individual pieces in a pile of cordwood (in Figure 4, the component axis are vertical instead of horizontal as in a cordwood stack). Interconnections between components are made using nickel ribbons spot welded to the component leads. These spot welds have proved to be considerably more reliable than soldered connections. Before spot welding the ribbon, the component leads may be threaded through holes in rigid plastic separators which run the length of the stack and serve to keep the components in place, as well as to prevent the welded ribbons from undesirably contacting the components. After the interconnections are made, the entire assembly is potted in a suitable plastic. With mini-weld, considerable space saving is achieved because the stack of components can be essentially a free form in which each component occupies a volume most appropriate to its own particular dimensions. Ideally, however, the components would all be the same length measured axially and they would all have leads of the same material in order to facilitate spot-welding the leads to the nickel ribbons.

Figure 5 shows an exploded view of a typical assembly used in the RCA micromodule approach. Each component is to be mounted on a ceramic wafer of the shape shown, approximately 0.3 inches square. Electrical connections are made by metallizing strips out to the notches at the wafer edges where solder connections are made to the vertical wires. These wires serve both as interconnection between components and as the mechanical structure of the assembly.

---

2. A. Brodzinsky, L. Anderson, E. Hurlburt, A. Shostak, V. Wanner, R. E. Wiley, Final Report of ONR Study Group on Microelectronics, ONR, Washington, D. C. June 1960
3. To be published

The micromodule program is extremely large having been funded to a considerable extent by the Signal Corps. Many companies are working with RCA and the Signal Corps on this method.

In Figure 6 we see a proposed assembly method which is achieving great favor in the industry; so much so that the EIA has already established standards for component dimensions. It is worth noting here that foreseeable assembly techniques should make possible a factor of 100 reduction in the volume of transistorized electronics in relation to the volume of today's equipment. This would be a component density on the order of 10^6 parts per cubic foot, and as we shall see in more detail later on, this component density may well be the limit that one can attain by any method of miniaturization.

In Figure 7, we see the thin film method for constructing electronic assemblies in which the individual components begin to lose their identity since the components are created simultaneously with the construction of the subassembly on the circuit wafer. The particular structure shown in Figure 7 is the method used by G. Selvin at Sylvania Systems Division. It has great practical advantages over other known methods in the ruggedness of its structure, the quality of its seals and its ready adaptibility to large assemblies. The resistors and capacitors in this method are to be formed by the deposition on ceramic substrates of various combinations of metallic and insulating films. Evaporation, sputtering, chemical decomposition of vapors, and many other means have been proposed as methods for depositing films.

One method of providing the active elements such as transistors and diodes for film circuits would be to mount the bare transistor or diode dice directly on the ceramic substrate and connect the electrodes of the transistors and diodes to the other circuit elements by thermocompressive bonding. Naturally, if bare transistor dice are to be mounted on the ceramic substrate, some form of hermetic seal must be provided over the entire substrate to protect the semiconductors. One proposed method is shown in Figure 7, namely, a glass hat, over the entire wafer, sealed to the wafer at the edges with a lower melting point glass. [3a]

From the point of view of pure physical assembly, it is probable that component packing densities can be achieved with thin film circuitry which are actually greater than can be effectively utilized in electronic systems. Nevertheless, these assemblies may come into common use if they can compete favorably with such methods as miniweld, and discrete components inset in the substrate, purely on the basis of cost, performance, and convenience.

Figure 8 shows a method of assembly in which the semiconductor device and the supporting substrate become one and the same, to a considerable extent. Circuit functions are performed more or less in the same way as in the methods of circuit structure discussed so far, but the Gordian knot is cut by simply eliminating the circuit structure as a separate item.

In Figure 8 the combination of elements shown is hypothetical and represents no practical circuit. It does show how a transistor, a capacitor, a "resistor" and a diode might be combined. It also indicates that no matter how clever one is he will need to connect some wires between regions of the semiconductor. Also, examination of the diode at the right will show one of the topological limitations of such assemblies; if one wished to connect certain elements in series rather than in parallel, rather complex layer structures may be required. In the

---

3a. G. J. Selvin

semiconductor art, the art of making crystals having three chemically different layers of controlled thickness has been mastered, and we are beginning to be able to handle four layers; more complex structures will be difficult, although not impossible by any means.

Miniaturization theoretically achievable by such means staggers the imagination. However, as we said before, the dissipation of heat from electronic assemblies places a practical limit on the possible miniaturization so that there is little point in packing circuits more densely than possible with the various discrete-component or thin film approaches.

An additional problem peculiar to integrated circuitry is the reduction of yield associated with the inseparability of the individual devices. The yield problem in semiconductor fabrication is serious now, even though the majority of the devices combine as inseparable, unitary assemblies, the more often we will have to throw away good junctions because one or more of the other junctions in the assembly may be bad. To be sure, the probabilities of junctions in an assembly being good or bad are not independent since the mere fact of simultaneous fabrication tends to result in junctions which have less than average variation among themselves. However, without some drastic change in method, fabrication yields will be severely reduced by the use of integrated structures.

The problem of poor yield can, in fact, be overcome by the same sort of redundant circuitry to be described later in this paper for solution of the reliability problem in any complex assembly. In fact if we consider failure during fabrication to be the same thing as failure after fabrication, we see that yield and reliability become one and the same. The problem of heat dissipation cannot so easily be overcome, however, unless we find new schemes for processing information.

Such schemes might not even be electronic. One can imagine mechanical, thermal or optical data processing with means provided for conversion between electrical and the other type of signal at input and output. The point here is that there is a wide open field for invention of new devices. In fact it has been suggested by some that the best approach to the problems of miniaturization is to plan to perform all electronic functions by new methods using new solid state phenomena. It has even been implied that it should be possible to find ways of performing any desired electronic system function by suitable modification of the interior of pieces of solid material, with a consequent elimination of the circuit as such. The names molecular-electronics or functional-block electronics have been coined for the proposed approach. Unfortunately, one cannot schedule such inventions. Advances of this sort come only from basic research and they come completely unpredictably; the only generalization one can make is that the laboratory with the best combination of basic research capability and practical orientation will be the most productive.

Figure 9 shows an example of a true functional device being worked on by the Sylvania Semiconductor Division's laboratories under the direction of Dr. T. Longo. It is an information-storage-shift register consisting of an array of PNPN switches all having in common two of their layers. Each switch can store one bit of information by being in either of two stable states, low or high impedance. The low-impedance or "on" state involves the build up of a large density of stored minority carriers in the common N-type layer. If only one switch is on, a lateral voltage pulse, applied as shown in Figure 7, can shift these stored minority carriers to the right one notch, thereby turning off the one switch and turning on the next switch to the right. A little thought will show how this can be used for shift-register action. That this is a functional device and not an integrated circuit can be seen from the fact that the shift register would not work if it were cut up into separate switches and then reconnected with wires, since there is no minority carrier conduction in metallic wires.

One can summarize the situation on integrated circuits and functional devices by saying that the day will probably not come when either of these methods is the dominant form of electronic assembly; sufficient miniaturization to reach the heat limit can be achieved with discrete-component and thin-film methods. On the other hand, there will be some places where integrated circuits will offer the system designer simplified assembly and greater convenience of use, with over-all lowering of costs. Thus, we may expect such assemblies to come into increasing use. If functional blocks are to be used to extend miniaturization beyond the 100 X reduction possible with the other methods discussed - and there seems to be very little reason for such an effort - the basic figure of merit on which the functional block should be judged is the amount of heat dissipated by the functional block in comparison to the dissipation of a standard assembly to perform the same job.

Figure 9 indicates that the transistor and the ferrite core are examples of functional devices, as is the newer R-C-line tuned amplifier of Westinghouse. The transistor and the core fill the bill since they are - or were, when invented - new ways of performing electronic functions by modifying the internal characteristics of pieces of solid material. Certainly the world is always waiting for new inventions; and if there is a real desire on the part of any organization to stimulate progress in functional electronics, the only way that desire can be fulfilled is to stimulate the atmosphere of free and basic research from which such inventions come.

Naturally a system with 10^6 parts has very little chance of operating at all if its function depends on the simultaneous operation of all its parts and if each of its parts has any reasonable probability of failure. However, there are coming to be excellent ways around this problem.

Redundancy is a fairly obvious solution and its benefits were analyzed several years ago by the late John VonNeuman. Recently, Professor Widrow and his associates at Stanford[4] have shown how to take full advantage of circuit redundancy. Their method is shown in Figure 10.

Clearly, the simplest thing to do is to have each operation performed multiply, and to decide the result by majority vote. Electrically, this is easy to accomplish if the two binary states are + and - 1 instead of 1 and 0. An algebraic sum-taker, indicated by Σ in the figure, followed by a threshold detector set at zero, takes the vote.

The sophistication added by the Stanford group is to add an analogue element - i.e., a simple volume control, shown in Figure 10 as $a_1 a_2 \ldots . a_p$ - after each of the multiplexed channels. The volume control is adjusted in accordance with the past performance of each channel so that the sum finally is weighted with the more reliable channels dominant.

The reliability of each channel is monitored in terms of its frequency of deviation from the majority decision and the proper volume control adjusted appropriately and automatically. Thus, the machine becomes an adaptive, or self-healing system.

Whereas the probability of proper operation of a chain of N elements can be represented as $P_{chain} = P^N_{elements}$, and becomes uselessly small as N gets large for any practical value of $P_{element}$, the probability of proper operation of a redundant adaptive system can easily be made greater than $P_{element}$, even for small degrees of redundancy.

---

4. Widrow, Pierce and Angell, Birth, Life and Death in Microelectronic Systems, IRE Transactions on Military Electronics, Vol. MIL5, pp. 191-201, July 1961.

In Figure 11 we see the excellent performance results that can be achieved using this adaptive system. The method of Widrow et al is indicated as the optimum-decision-element curve. It shows, for example, that redundance of a factor of 10 can make a 10^{10} improvement in system performance. The results achievable with the earlier proposals of Moore and Shannon[5] and VonNeuman[6] are shown for comparison.

Thus, reliability will cease to be a component manufacturer's problem when the component manufacturer has done all he can - particularly when he has eliminated all sources of systematic error in which environment or time affects all components similarly - and will become a system design problem. There should thus be no barrier to the creation of systems having any number of components, and if other problems could be overcome, such systems could be indefinitely small. Unfortunately, these "other problems" may indeed prove insoluble and perhaps this is the main point of this paper.

Briefly put, heat and its disposal will certainly set a lower limit on the useful size of any electronic system. In fact, if we think of logic networks using active PN junctions, resistors, capacitors and inductors in familiar ways, we conclude that systems will have to get rid of roughly 10 mw per component. Thus, a 10^6 component system will dissipate 10 kw.

The problem of minimizing the heat dissipation in conventional computer circuitry has been given all too little study up to now, but both experience and the theory that has been undertaken confirm the figure I have quoted.

The best work in this field that has come to my attention is the work of J. J. Suran, who has studied the effects of operating frequency, component tolerance, power supply tolerance, and required systems reliability, on the minimum required operating power of conventional transistor flip-flops.

In Figure 12 we see Suran's[7] results for a family of flip-flops designed optimally to operate at various frequencies. The transistors used all had the same f_{max}, but the circuits for the different frequencies were not the same, of course. These are experimental data, but theoretical calculation predicts essentially the same result. The problems here are basic and related to the well-known fact that non-resonant charging of a capacitor requires a total of twice the energy stored in the capacitor, regardless of the value of the series charging resistance.

The use of PN junctions and wiring of lower capacitance reduces the power, of course, and straight-forward linear shrinkage of all dimensions will reduce capacitance; thus, the energy expended per bit of information will go down with circuit size. However, as we indicated earlier, it is the wiring, not the semiconductor devices, which will be reduced in size, so no great improvement in junction capacitance can be expected. Furthermore, linear reduction in dimension gives linear reduction in wiring capacitance but cubic reduction in volume. Thus, power per unit volume expended in charging and discharging wiring capacitance will rise inversely as the square of linear dimension.

---

5. Shannon and Moore, "Reliable Circuits Using Less Reliable Relays," J. Franklin Institute, 262, pp. 191-208 and 281-297, September and October 1956.
6. J. Von Neuman, "Probabilistics Logics and the Synthesis of Reliable Organisms from Unreliable Components," Automata Studies, Princeton Univ. Press, 1956.
7. J. J. Suran, "Circuit Considerations Relating to Microelectronics," Proc. IRE, Vol. 49, pp. 420-426, February 1961.

Of course, somewhat faster devices could be used to minimize the power consumption at any given frequency of operation, but that would correspond to operation on that portion of the curve of Figure 12 far below the device capability. Since system complexity and speed can be traded linearly, the important thing being the number of possible binary decisions per second, it is not likely that the active devices in a miniaturized electronic system will be operated far below their best frequency capability. Faster devices than those considered on this chart will give essentially the same result for power when used at the same relative frequency.

Naturally, a computer system operating with smaller signal levels will dissipate less power. Smaller levels can be tolerated when more precise components are used, but unfortunately high precision components become harder to make the smaller they get. Suran has also considered this problem.[7]

In Figure 13 we see his results for the required operating power of flip-flops in relation to resistor and power-supply tolerance when the signal level is made as small as possible for reliable operation. The so-called worst case design method is used.

In the figure we see the power per flip-flop plotted as a function of resistor tolerance with power-supply tolerance Δ E as a parameter. Thus, if we use 20% resistors and allow the power supply to vary 10%, the circuit power will be 200 mw. The curve marked equitolerance locus shows the circuit power if Δ E and the resistor tolerance are taken to be equal. Here again, a power of a few milliwatts per component is the minimum practical value if we assume that the flip-flops in question had ten or twenty components each.

At the outset it was stated that a component packing density of 1000 per cubic inch (or 10^6 per cubic foot, more or less) was a practical limit on miniaturization, and we should now be in a position to justify that.

Suppose we wish to build a system of 10^6 components, each dissipating 10 mw. This comes to 10 kw, as we have said. Let us see in a roundabout way what is a sensible size for such a system. First, how big would the power supply be? For continuous operation, certainly larger than one cubic foot if we include the heat engine to drive it. For a few minutes of operation probably a battery source could be smaller, but the real limit is in the size of the cooling apparatus.

Systems of the type we are considering would surely be mobile (truck-, air-, satellite- or missile-borne) and tap water would not be available as a cooling agent. However, in one way or another, the heat would have to be dumped. Fluid cooling would certainly be necessary for such a compact heat source and a fluid temperature of 100°C is probably maximum. The heat could be dumped in an air-fluid heat exchanger (like the so-called radiator of an automobile) or it could be transferred via a heat pump to a true radiator at a higher temperature for space use.

Air fluid exchangers of minimum size range between half and four cubic feet to handle 10 kw at a fluid temperature of 100°C, depending on whether one includes the fan and motor to move the air. If one used the ram air in the case of an airplane installation, there is the hidden factor of the increase in main power plant size required to overcome the drag of the heat exchanger. These figures are derived from data supplied by the manufacturers of heat exchangers.

For space use, we can calculate the area of the radiant surface needed, provided we include the power to operate the heat pump. If the cooling fluid is at 100°C and the radiant surface at a dull red heat, at the theoretical efficiency the heat

pump will consume an additional 15 kw. To radiate 25 kw at a dull red heat, a perfectly black surface requires 7 square feet according to the Stefan Boltzman law. These calculations are summarized in Figure 14. The first equation is the relation between total power to be radiated and the power dissipated in the electronic system at 100°C, taking the efficiency of the heat pump to be the theoretical Carnot value. The second equation is simply the Stefan-Boltzman law and shows how rapidly the size of the radiator can be reduced as the temperature is allowed to climb, in spite of the fact that more power must be expended in the heat pump at higher temperatures. The final equations show that 10 kw dissipated at a system temperature of 100°C (373°K) requires 7 ft^2 of radiator at a dull red heat (900°K). Incidentally, we can also see from the second equation that to radiate the power from a radiator at the same temperature as the system would require about 100 ft^2, a size that seems impractical.

Thus, no matter how one looks at it, there appears to be no point in going to less than 1 cu. ft. for 10^6 components. This density should be achievable with discrete components, since it allows an average space of 0.12 x 0.12 x 0.12 inch for each component. Surely, if the active part of a transistor is no larger than 10^{-8} or 10^{-9} in^3, the wiring and environmental protection should not require more than 10^{-3} in^3, 10^5 to 10^6 times more. Dr. Suran[7] has analyzed the problem of heat flow in electronic assemblies in more detail with fewer restrictive assumptions and his conclusions are somewhat similar.

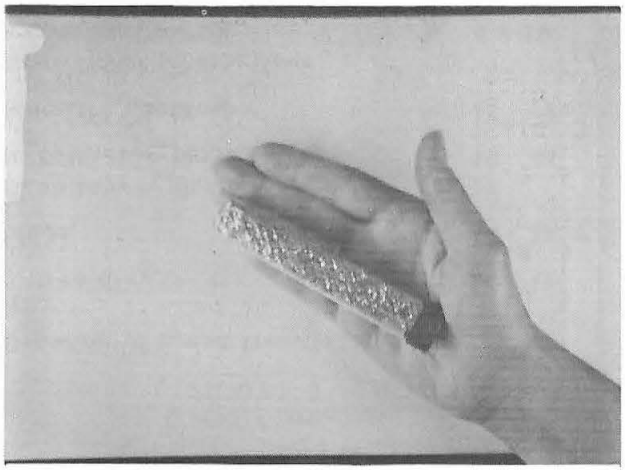


FIGURE I

Vial containing 250,000 complete "transistors", i.e. mesa transistor dice.

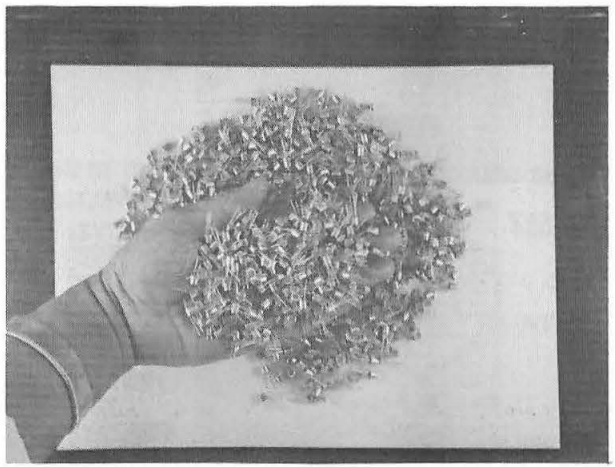


FIGURE 2

1000 packaged mesa transistors, each with its separate glass-to-metal enclosure and with electrical leads, ready for installation into circuits.

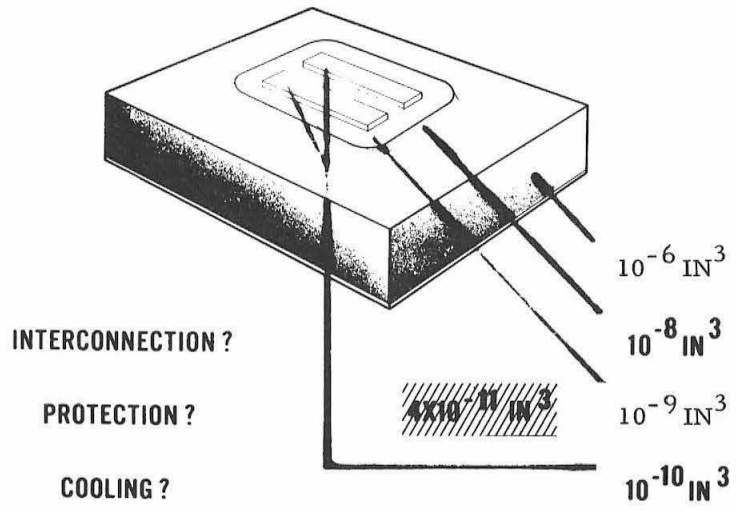


FIGURE 3

Drawing of mesa transistor die.

Entire die,	10^{-6} in^3
Mesa	10^{-8} in^3
Base layer, down to collector junction	10^{-9} in^3
Metallic contacts	10^{-10} in^3
Base layer of smallest known unit	4×10^{-11} in^3

Interconnection, protection and cooling provisions take up most of the volume in electronic assemblies.

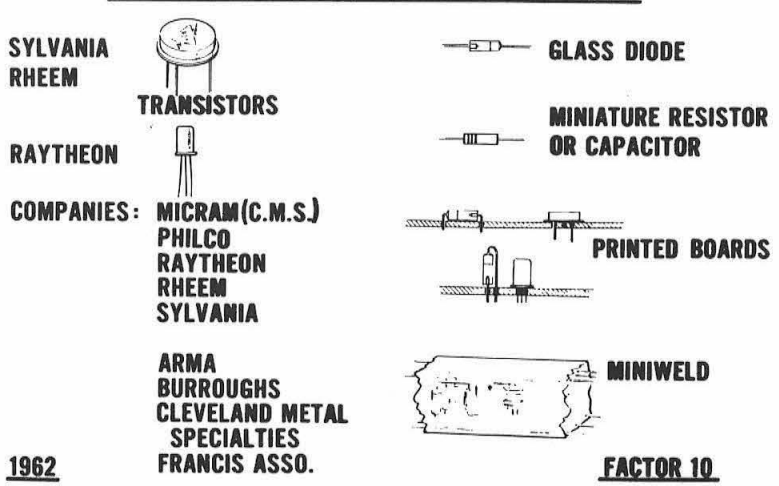
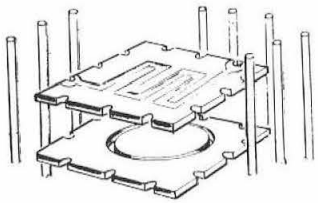


FIGURE 4

DISCRETE COMPONENTS - MICROMODULE
(ONE PER SUBSTRATE)



R.C.A. AND SUBCONTRACTORS

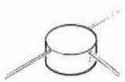
1963 FACTOR 20

Figure 5

DISCRETE COMPONENTS-DOT PACKAGES
(INSET IN SUBSTRATE)

COMPANIES

HUGHES
MALLORY
PACIFIC SEMICONDUCTORS INC.
TEXAS INSTRUMENTS
TRANSITRON
SYLVANIA



ARMA
DIAMOND ORDNANCE FUZE LAB.
HUGHES
ETC.

1963 FACTOR 100

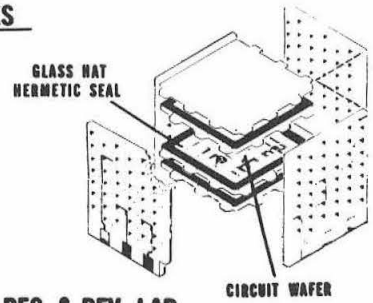
Figure 6

FILM CIRCUITS

MANY COMPONENTS ON COMMON CERAMIC SUBSTRATE, INTERCONNECTIONS BUILT IN

COMPANIES

BELL TEL. LABS.
I. B. M.
INTELLUX
R. C. A.
SPRAGUE
SYLVANIA
U. S. ARMY SIGNAL RES. & DEV. LAB.
VARO MFG.



1965 FACTOR — 500

Figure 7

This particular method shown is that of G. Selvin of Sylvania Systems Division.

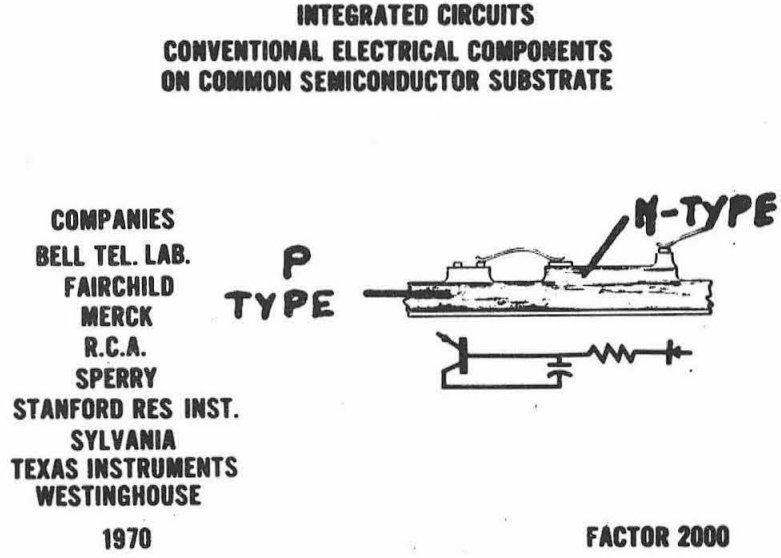


Figure 8

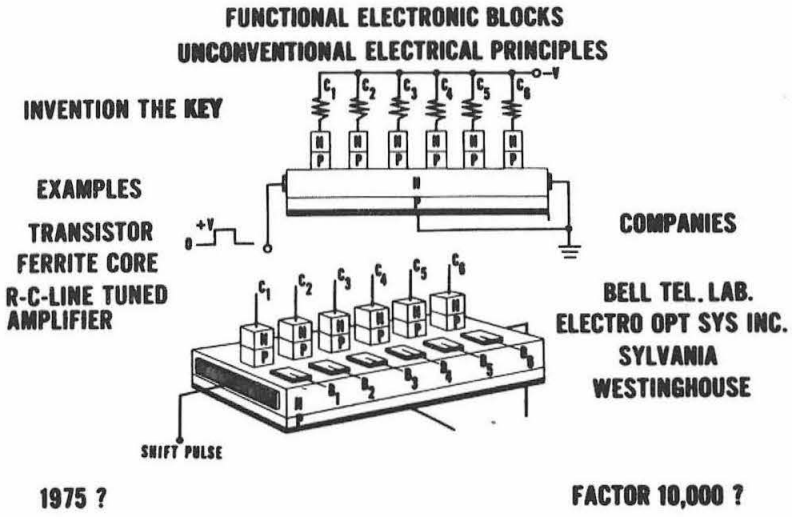


Figure 9

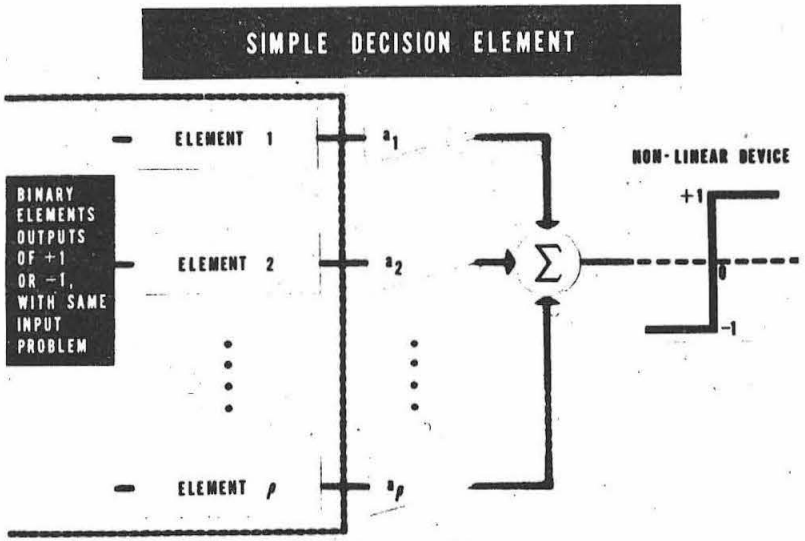


Figure 10

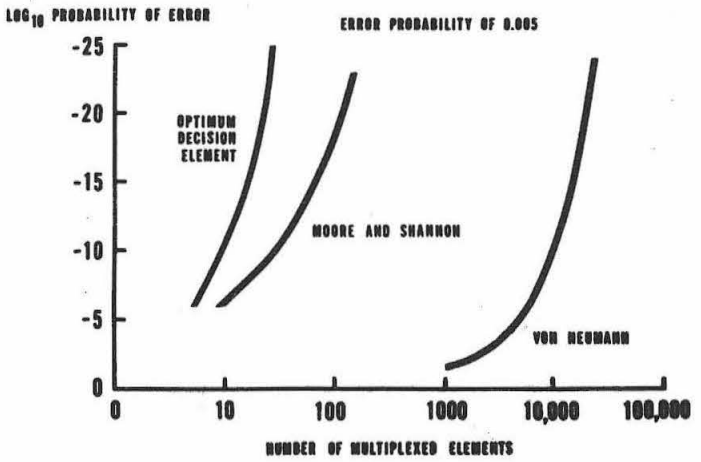


Figure 11

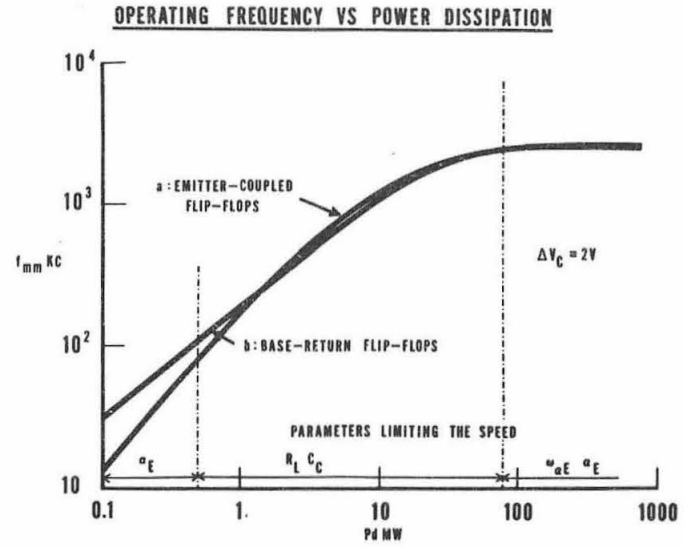


Figure 12

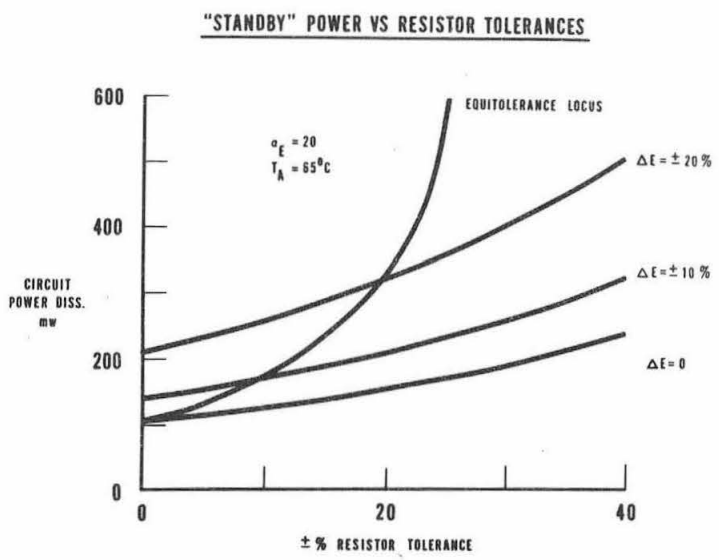


Figure 13

Figure 14

Calculation of size of radiator needed to dissipate 10 kw in free space.

$$P_{rad} = P_{sys} \frac{T_{rad}}{T_{sys}}$$

$$A_{rad} = \frac{P_{rad}}{K T_{rad}^4} = \frac{P_{sys}}{K T_{sys} T_{rad}^3}$$

$K = 5.3 \times 10^{-9}$ WATT / FT^2 DEG^4

$T_{sys} = 373°$ KELVIN

$T_{rad} = 900°$ KELVIN

$P_{sys} = 10^4$ WATT

$A = 7$ FT^2

COMPUTERS

Gardiner Tucker
International Business Machines Corporation

The three fields - Data Processing, Communications, and Instrumentation, are beginning to coalesce, and it will be harder to differentiate them usefully in the next decade than it has been in the past decade. The reason is two-fold: as we become more sophisticated in each discipline, we find an increasing need to pull in techniques from the other two; the acquisition of information, its gathering and dissemination and its manipulation are all required to perform a useful service for society; and as we progressively remove human links between these functions, the corresponding disciplines must interact.

Let me cite some examples of the interaction of these fields. In instrumentation, there is increasing concern with making the information acquired not only as accurate as possible, but also as directly useable as possible. Accordingly, the raw information is transformed to linearize, digitalize, shift zeros, compensate for drift or for hysteresis. This requires the use of analog and digital techniques borrowed from the field of Data Processing. Telemetry, e.g., of data acquired by satellite, requires intimate cooperation of instrumentation and communication techniques. A modern communication system, with its ability to select and test channels, and adjust codes and rates according to channel characteristics, and its message storage and assembly and selective addressing features is beginning to resemble a computer strikingly. As data processing systems move from record keeping functions into planning, monitoring and control functions in industry, increasingly information about the current state of affairs must be entered into the system automatically at the point of origin and the results of processing must be made continuously available at the point of influence. This requires instrumentation techniques at the terminals and communication techniques throughout.

One of the great technical challenges facing society has been the challenge of producing enough energy and of making it readily available when and where it is needed so as to liberate man from the burden of physical drudgery. The scientific basis for meeting this challenge was established during World War II with the development of nuclear energy. There remain tremendous engineering problems of controlling this energy and distributing it and making it economical. An analogous technical challenge is that of producing information and of making it readily available when and where it is needed so as to liberate man from the burden of mental drudgery and create new opportunities for progress. The last decade has seen tremendous progress in the creation of raw concentrated data processing power. We are faced with two major engineering challenges for the next decade: to increase computer power much further; to make this computer power readily and economically accessible when and where it is needed.

Each of these challenges has implications for device technology. Let us first consider the drive towards greater computer speed and power. The achievement of the speed and concentrated power of a STRETCH type machine has advanced technological and systems design as its name implies. It required the development of very high speed transistor devices and circuits and of very high speed cores, and of many tricks of organization. To go substantially further introduces new kinds of problems. If all the transistors and cores in STRETCH could be replaced with infinitely fast devices, we know that the computer would not be infinitely fast. There is the very fundamental limitation that information cannot be transmitted faster than the speed of light or 30 cm.

per nanosecond. If future computers are organized as are those of today, with separated boxes for memory, logic, arithmetic, etc., then there is maybe little point in making individual devices faster, for we may spend all our time in waiting for information. The machine calls for an instruction from memory and waits while it reaches a register. After the instruction is decoded, it calls for data from memory and waits. When that arrives, it is modified and sent back. All the time is spent in internal communication. But, you say, let us organize the solution of our problem in such a way that we can anticipate several steps ahead what we shall require from memory and start it on its way in time. Unfortunately, this gives greater speed for only a very restricted range of applications, because the great power of a computer lies precisely in its ability to change its internal operation abruptly depending upon the state it has achieved. The internal operation of a computer is carried out by the continual flow of data from memory to logic to arithmetic to memory, with transformations performed at these locations. Now that the transmission times are becoming comparable with device switching times, transmission time must also be minimized. This means that transformation and storage elements must be located together to minimize data paths. The implication for device technology is that there will be less interest in seeking optimum devices for each separate function. For if they are technologically incompatible with one another, then the time space and money required to couple them together in this integrated way defeats their virtue. The emphasis must be on a consistent technology for memory, for logic and for their connection.

The point may be illustrated by consideration of a modern high speed core memory, for example from STRETCH. Classically, it is conceived of as a stack of core planes. However, that is not what it looks like. It is so surrounded with drivers, sensors, amplifiers, control circuits, and by wires swarming everywhere, that the cores are quite lost. Imagine attempting to mix transistor logic intimately into such a memory.

An example of a consistent technology is cryogenics. By exploiting the same materials and physical phenomena, it may be possible to perform storage, logic, amplification, shaping, and connection. This permits tremendous design flexibility so that logic rather than devices may dictate machine organization. From this point of view, we have been disappointed in transistors. Although they have permitted smaller computers with low power consumption, and longer life, they have not facilitated revolution in computer organization to achieve major increases in processing power.

Let us now turn to the second challenge for the decade: to make information processing readily and economically available. There are two approaches. One is to give distributed access to a larger central computer so many jobs are performed quasi-simultaneously and each user employs a small slice of big machine time through a local console. The other is to make small, economical but powerful separate machines. The first approach requires that the computer be placed in the center of a complex communication network with Input and Output gear at its terminals. The problems of the network are shared with the communications industry and their device implications will be discussed by Mr. Pierce. The major problem of a pure Data Processing nature is the methodological or programming problem of handling a multiplicity of applications on a computer at the same time in such a manner that each terminal appears to have an entire computer under its control. The device implications of this horrendous problem are not yet clear.

To progress substantially towards economical but powerful computers, the luxury of individual devices separately fabricated, separately encapsulated or mounted, then incorporated into separate circuits or assemblies, then connected together to form functional units, can no longer be tolerated. The cumulative costs of these successive stages, each with its own processes, makes the resultant equipment too costly to serve effectively large areas of need. Even if individual cores and transistors were free,

it would not reduce the cost of computers dramatically. What is needed is technology or technologies for devices and their interconnections which permits the fabrication together in coherent processing steps of entire functional units. The goal is a consistent technology with emphasis on manufacturability.

The implications of the Data Processing field for solid state devices come from a consideration of the entire organization of a computer. More important than any characteristic of an individual device such as its speed or size or cost is the requirement for compatible technologies throughout a machine so that the various individual elements may be fabricated together and may function together. These are requisites to meeting the challenges of power and economy which are generated by the needs of society in the coming decade.

The foregoing extemporaneous remarks were delivered by Dr. Tucker who substituted for Dr. E. R. Piore.

COMMUNICATION

John R. Pierce
Bell Telephone Laboratories, Inc.

The glory and importance of science and technology lie in the particular, in the concrete example. We sometimes despair whether philosophical, ethical and political considerations which in principle embrace all of the universe, or at least all of man's society, will move that society or its members one step forward. Yet one particular device, the telephone, affects each of our lives profoundly, and it has called a great industry into being. We all see the effects of the automobile and the atom bomb, both on social and political institutions and on individual men. Such influences operate on a smaller, more private scale as well. The successful functioning of optical masers has brought about radical and quick changes in the lives of many first rate, independent technical men, changes which could not possibly have been accomplished by executive action.

It is clear that while broad goals and a sympathetic environment may smooth the path for progress, progress itself, whether in science or technology, is the result of concrete, particular advances in understanding and application. This colloquium itself is a tribute to the power of new understanding and new devices to change the worlds of science and technology, popularizing solid state physics and revolutionizing electronics.

Earlier speakers have had the fun of talking about the particular devices which have worked and are working a revolution. I am faced with the less substantial and perhaps thankless job of discussing the future application of solid state devices in communication. One course might be to cite statistics concerning the rise in the use of solid state devices and to make estimates of future needs and uses. I am no good at this, and I really doubt if such figures would be very helpful. We do not need figures to tell us about the ubiquity of transistor radios and hearing aids, or the millions on millions of magnetic cores at large in computers. The present speaks for itself, and there are not any meaningful figures about the future possibilities of solid state devices.

Thus, what I have to say is both nonquantitative and general, and I could not defend myself if I were accused of talking like a philosopher rather than a scientist.

Quite aside from the ingenuity or beauty of solid state devices, I propose to ask, what will solid state devices do that will change our world by improving communication?

Solid state devices have several revolutionary advantages. Foremost among these are low-power consumption, long life and low cost. In some cases the long life and low cost are merely a potential advantage, but they will be more universally attained. I believe that the most revolutionary effects of solid state devices will stem from these three properties, which make a broad use of complex equipment technically and economically feasible.

Solid state devices will have revolutionary effects other than those dependent on low power consumption, long life and low cost. A three-level traveling-wave maser is not cheap, and the cryostat required for its operation is not low power. Yet, the use of such a maser in the ground receiver of a satellite communication system could mean a satellite-borne transmitter with a power of 1 watt rather than 10 watts. Because weight increases with power and, indeed, is nearly proportional to power,

the existence and use of the solid state maser means that a given launching vehicle and launching cost can orbit up to 10 times as many transmitters as it otherwise could. This is certainly a revolutionary impact of solid state on the new, promising but difficult art of satellite communication across oceans.

Similarly, optical masers are neither cheap nor do they have a very low power consumption, yet their existence will revolutionize research aimed at very broad-band communication. At last it seems possible to extend the coherent signals and techniques of modulation and amplification characteristic of electrical communication to the optical range of frequencies, a range providing thousands of times the bandwidth heretofore available. Certainly, infrared and visible light could travel for miles in special atmospheres confined to pipes. For many years the communication revolution presaged by optical masers will be a revolution of research, but a revolution of application and exploitation seems sure to follow. In the light of our present knowledge we cannot say when or exactly how.

Solar cells are not cheap today, but they provide the only long-enduring power supplies used in space vehicles, and it appears that they will continue to provide the most long-lived and reliable power supplies for moderate powers. Surely this is revolutionary. I am sure that there must be other examples of revolutionary impacts of solid state devices on communication which are not necessarily linked with those three qualities I cited earlier: low power consumption, long life and low cost.

We come closer to what I have in mind in considering an experimental low-power microwave system which has been in continuous operation at Holmdel, New Jersey, for over three years. Because this system uses vacuum tubes as well as transistors, the power drain of the repeater is close to a hundred watts. Because the potential long life of transistors has not yet been universally attained, and because we must still use vacuum tubes to obtain substantial microwave powers, five transistors and several microwave tubes have failed during the 3-year test period. The goal of a very-low-power repeater capable of trouble-free unattended operation for years has not been quite attained, but it seems very close.

As long as microwave repeaters require large powers and appreciable maintenance, they will be expensive. As long as they are expensive, it will be desirable to span a given route with as few as possible. This will lead to tall, expensive towers, high expensive sites, and expensive surveying and engineering. Further steps based on solid state art could make microwave repeaters so cheap and long-lived that it would be most economical to use many with low towers at easily accessible roadside locations in getting from here to there. Here low-power consumption and long life are especially important, and low cost is highly desirable.

Let us think also of the network of wires which supplies us with telephone communication. If we examine a pair of wires leading to a subscriber, we will usually find it idle. Even when the pair of wires is in use, it is carrying only one telephone conversation.

If we examine a pair of wires interconnecting distant cities, the situation is otherwise. Such pairs of wires are almost always in use. Carrier telephony makes it possible to send two dozen speech signals over two pairs of wires, and a number of the channels provided are ordinarily in use. Coaxial cables and microwave radio relay systems provide thousands of channels at lower costs per individual channel.

Electronics has served us well in long-distance transmission. We load our wires heavily, and by means of switching we make them available to many subscribers. This is the economical thing to do, for while electronic equipment gets cheaper,

copper gets more expensive, and we should make the most of it.

The savings which electronics have made in long-distance communication are great and important. Perhaps it is partly because of such savings that 75% of the telephone investment is the local plant in cities and towns. This local plant includes wires and switching for local transmission, and the telephones themselves. How are we to use electronics to save here?

Suppose that we apply the conventional techniques of carrier telephony to the local plant even in a very restricted way, using it just to make the usage of wires linking central offices more efficient. Central offices become overloaded with both heat and equipment, and the power bills rise alarmingly. Clearly, such techniques are not practical for wires leading to small groups of subscribers.

Suppose that we try to share wires among more subscribers. This can be done by putting electromechanical switching devices called concentrators on telephone poles. While this can be profitable, I doubt if anyone regards it as a final solution.

The problem of the efficient use of the wires in the local telephone plant is typical of the sort of communication problem that solid state devices can solve.

An experimental solid state system for sending 24 telephone channels over two pairs of wires has already been tested in the Bell System, and is now being developed for use. This is a pulse code modulation system. Speech signals are converted into off-on pulses, and a million and a half per second are sent on a pair of wires. Repeaters located in telephone manholes amplify and reshape the pulses every mile. The small power required by the repeaters is carried by the same wires that carry the signals.

In this first step on a road of the future, low-power consumption is essential in order that economical operation be achieved at a remote and isolated location. Long life is essential in order to avoid excessive maintenance costs. Of course, a low or at least a reasonable cost is necessary to make terminal equipment of such a system cheap enough to justify use over the short distances encountered in local plant, and to make a repeater every mile economical.

This is one first step in exploiting the advantages of solid state devices in bringing electronics to new locations and new uses. We see that this step involves the unattended operation of complex but low-power equipment at remote locations. A next reasonable step would be to put substantial parts of telephone switching systems at remote locations, so that efficient multichannel transmission circuits could be connected to subscribers close to their telephones, shortening lightly loaded, seldom used, expensive individual wire circuits. An experimental switching system of this sort, called Essex, has been built and tested in the research department of the Bell Laboratories.

Once one accepts the installation of complicated, low-power, economical, electronic equipment at remote locations, and convinces himself that sufficiently long life can be realized so that maintenance at such remote locations is supportable, many other things become possible.

Already in a few instances in the Bell System, switching equipment is located in industrial areas and the operators' switchboards associated with it are located in more pleasant business districts some distance away. Ultimately, compact electronic equipment may be placed as a matter of course in unmanned locations and all service changes, including assignments and reassignments of numbers, may be carried out from remote locations over electrical communication channels.

Already pocket transistor paging receivers summon subscribers to the nearest telephone in response to a call. Mobile telephony, however, is confined to automobiles, and is not in very widespread use. Vacuum-tube radio transmitters and receivers give very valuable but rather primitive service to some thousands of auto owners. Special generators and batteries are required. A subscriber has access to only one crowded channel, or at best, a very few crowded channels. If this channel is not frequently busy, it is frequently unoccupied. Usually it is both.

High-quality mobile telephony calls for complicated subscriber equipment - equipment which the central office can direct to one of many available frequencies. Vacuum-tube equipment would be impractical, but low-power drain, long life, low-cost solid state equipment may be practical. If frequencies for mobile telephony are made available, solid state devices may give us mobile telephony of a quality comparable to other telephone service and of considerably reduced cost. Telephones in cars could become as nearly universal as car radios.

But voice communication is not all of communication. Despite the alternative facilities provided by teletypewriter, we still send and receive letters, carrying paper from here to there for the sake of the symbols written on it. Can we not communicate more or even most of our written words, or at least our typed words, electrically?

To do this requires on the subscriber's premises equipment which, were it not for solid state devices, would appear unduly complicated and fallible in comparison with the telephone set. It also requires a means for reducing material to a convenient machine-readable form and for printing it out - a machine comparable in size, speed, flexibility, convenience and cost to an electric typewriter. Solid state devices may make such a machine possible.

Who knows what other useful equipment the man of the future may find in his home or office. Facsimile? Television for conferences or even for person-to-person use? With advances of this sort we may expect an increased range and flexibility in our use of communication, in obtaining, in tracing and retrieving information, in making reservations, purchases and payments, and in conferring at a distance rather than in traveling to confer. Solid state enciphering devices will help to make this practical.

I think I have said enough concerning where solid state can and, I am sure, will eventually take us in communication. It will provide spectacular special functions or savings. The maser as a ground receiver can greatly reduce the power required in communication satellites, and hence the cost of satellite communication. Optical masers promise us the extension of communication techniques into an entirely new and vastly great frequency range. Solar cells provide the power for satellite communication.

However, an even greater revolution depends on the low-power drain, long life, and low cost which solid state devices can attain. These properties will make it technically and economically feasible to put complicated transmission, switching and terminal equipment at remote unmanned and at subscriber locations, and to control it remotely. This will result in diffuse communication systems which are far more complicated and highly organized than today's more centralized systems. Such systems will be more economical of copper and of bulky facilities, including expensive sites and structures, than today's systems are. They will give more varied and more flexible service, including mobile communication and the transmission of text and pictures as well as voice, and the application of such transmission to a variety of needs, including conferences, reservations, purchases, payments, and information retrieval.

How do we attain this utopia? Many obstacles we will see only when we stumble against them. Right now, however, I see a number of roadblocks which I wish we could overcome.

In this communication wonderland of the future, the life of components must be far beyond what we can put up with today. Presumably the repairman will ordinarily confine his operations to replacing complex plug-in units, but with apparatus scattered far and wide we cannot afford much of this. Will this favor magnetic units at the expense of semiconductors? I do not know, but near-eternal and stable life is absolutely essential.

Economies in picture transmission (if they are to be attained) and many other desirable functions call for very cheap, compact, long life, low-cost storage of millions of bits. Random access is not necessary. I wish I saw some solution to this problem.

Complete electrical control and operation of switching centers call for large, cheap memories with electrical write-in and nonerasable readout. Magnetic drums or discs are quite good, but some would like faster operation and random access. I hope that this problem will be solved.

I also hope that someone will develop a simple, inexpensive, flexible typewriter suitable for general secretarial use, which will produce a satisfactory machine-readable record as well as typescript, and which will print out machine-readable records. Such a device is essential to routine and universal electrical transmission of text and to any routine and universal partnership between man and machine in handling man's vast and perplexing written records.

INSTRUMENTATION

Bernard M. Oliver
Hewlett-Packard Company

Instrumentation is the branch of our modern technology which provides science with the tools for its research, provides engineering with the tools for product development, and industry with the tools for checking its production. In electronics these tools are oscillators, signal generators, voltmeters, oscilloscopes, counters, pulse generators, and so on - in other words, all the devices with which we stimulate other electrical equipment and measure its response. In fields other than electronics per se instrumentation provides such things as cardiographs, balances, pyrometers, spectroscopes, calorimeters, tachometers, microscopes and telescopes. In short, it provides the measuring and sensing organs of science and technology.

Recently instrumentation has taken a new turn. In some areas, such as missile test stands and satellite tracking stations, electronic instrumentation has assumed prodigious proportions, and is represented by complex systems for taping and processing data at an enormous rate.

I have been asked to talk about the future of solid state devices in instrumentation. I shall talk primarily about electronic instrumentation, because that is the field with which I am most familiar. Furthermore, the solid state devices, being themselves electronic, are certainly going to be applied there first. However, some of the qualities that these newest devices bring to electronic instruments will tend to qualify these instruments for use in other areas as well; and so we will have occasion to say a few words about some of these other areas.

Like anyone with a scientific background, I always feel a little uneasy when I am asked to make predictions. Scientific people like to make definite and accurate statements, and anything said about the future must be either indefinite or inaccurate. However, there is a method of prediction which is scientifically respectable nowadays. This is to survey the past and the present and extrapolate any trends which may be apparent. This is the way the weatherman works, the way ephemeridaes are calculated, and the way a fire control computer operates. I shall try to use this type of prediction rather than the Hugo Gernsback variety, stimulating though the latter may be to the imagination. I may go on a few flights of fancy; if so, please forgive me. I am not prepared to defend anything I say nor even my right to say it. I am fully aware that anything I may say may be knocked into a cocked hat by the sudden appearance, like a mutation, of some new discovery or other. With these apologies, then, let us proceed.

First, let us ask: What changes have already occurred in instrumentation due to the introduction of solid state devices? For one thing, the size reduction which everyone predicted would happen is indeed beginning to occur. This size reduction, incidentally, is made possible not primarily because the transistor is smaller than a tube, though this helps, but rather because it consumes less power. With less heat to dissipate, components may be packed together far more densely than ever before. Further, because less power is used, some of the components themselves, such as power transformers, tend to become reduced in size. This helps, too. However, in other cases the transistor has created component problems. For example, in many circuits operating at audio or fairly low frequencies, much larger coupling capacitors are required with transistors than with tubes. These capacitors, it is true, operate at lower voltages, but extensive development in miniaturization has been necessary before complete advantage in size reduction could be obtained.

The dense packing of components has led to new methods of construction, with which I am sure most of you are familiar. In the old days, in old instruments, we used to mount the parts on a metal chassis and simply wire them together, and the fact that they were far apart helped get rid of the heat. Today, as in computers, the trend is largely toward modular construction in instruments, with banks of components on plug-in printed circuit boards. It has taken time to develop the new components, and the new methods of construction and instruments using these new methods are just beginning to appear on the market. So I predict that electronic instruments, particularly the larger ones, will grow much smaller within the next few years. In fact, this year the Hewlett-Packard Company is replacing many of its older instruments with ones that are 1/4 to 1/10 their former size, as I am sure other concerns are also doing. Many of these new instruments will have improved performance and include new functions in spite of the smaller size.

Actually, there are cases where the size is dictated not by the circuits, but rather by the fact that the panel has to be big enough to hold the knobs, or that the meters have to be big enough to be read. Thus, for many simple instruments, size is dictated by the human hand and the human eye rather than by the contents of the instrument. It is because this is not true with the larger instruments that size reduction will be more dramatic in their case.

Compactness and lightness of course, make for greater convenience. Where formerly an experimenter might be all but buried by a mass of hot instruments surrounding him, he will now be able to place many more transistorized instruments within convenient reach and still have plenty of room to spare. Further, the instrument which might have filled several racks now can be designed to fit conveniently on his bench. Thus, miniaturization will place new classes of instruments at his convenient disposal in the laboratory.

I think that the reliability long promised for the transistor is beginning to materialize. This has spurred component makers to improve the reliability of other circuit parts. We perhaps have not reached the stage of reliability that John Pierce would like to see for unattended repeaters and some other applications, but things are much better these days than they were just a few years ago. Further, the lower circuit operating temperatures, made possible by the lower power consumption of transistors, tends to increase the life of all components in the circuit. Our experience indicates that with well-designed transistor instruments one should expect about an order of magnitude of improved reliability. We shall come back to this point a little later because it is important.

I would like to discuss next another aspect of solid state devices in relation to instrumentation. We tend to look upon solid state devices as simply replacements for tubes. However, it is a fact that many of these solid state devices do things that a vacuum tube cannot do. They have brand new properties associated with them, and sometimes these properties are very valuable. As a result, we can accomplish new functions in some cases, and in other cases do things just as well with fewer components. Let me cite some examples.

First, I would like to remind you that the transistor itself is available in both polarities - PNP and NPN. This adds a degree of freedom to circuit design that we did not have with vacuum tubes. It is as if someone had invented a vacuum tube using positrons so that one could connect it upside down in the circuits. Having both polarities available has led to simpler class B amplifiers and DC amplifiers in which potential does not accumulate through successive stages. Also, a saturated transistor, unlike a vacuum tube, carries current quite well in both directions. So, it makes an admirable bilateral switch.

We have tended to look for fancier solid state devices as time goes on, and have perhaps tended to overlook some of the potentialities of the simpler devices. I would like to call your attention to the fact that the humble diode in its solid state form has risen to new heights of sophistication, which in some cases, rival those of the transistor. A single solid state diode can now perform functions which were previously either impossible or else required many tubes and associated components. Let us look at some of these.

The silicon power diode is an example of simply improved performance attained through the solid state device. Here we have extremely high current capacity now possible (on the order of 25-50 amperes), together with high rectification efficiency. The use of silicon rectifiers in instrument power supplies has itself led to a considerable reduction of heat, even when we have had to use vacuum tubes for some of the other functions.

The photodiode is finding some application. This is simply a PN junction rectifier and, when reverse biased, it conducts current due to the generation of hole-electron pairs by the photons. It is a very cheap and good photocell and we are finding some applications for it in our instruments.

Another diode that is perhaps not so well known is the PIN diode consisting of a P-layer and an N-layer with an undoped intrinsic layer between. Such a diode has a very low capacitance because the conducting areas, the N area and the P area, are separated by the insulating I-layer in between. So it has in its non-conducting, or back-biased state, a quite low capacity. On the other hand, when one forward biases this unit, the carriers injected into the intrinsic layer render it conducting. Consequently, if one has simultaneously present, a direct current applied to forward-bias the unit and an RF voltage present across it, one can vary the conductivity of this unit to radio-frequency. Although the radio-frequency injects carriers on one-half cycle, it also takes them back out again on the next half cycle. Modulation of the conductance by the RF itself, is negligible when the oscillations are so rapid compared with the lifetime. Hence, a PIN diode can control a shunt conductance across a RF transmission circuit quite nicely with a small DC current. We find this to be a very useful device as an attenuator in wave guide for signal generators. It replaces much more cumbersome equipment formerly used and it can be electronically controlled.

An ordinary PN junction has the property that the junction capacitance varies with the back-bias. Carriers which extend up to the junction in the case of no back-bias recede from the junction as back-bias is applied, and the effect is as if the plates of a capacitor were being pulled apart. The more back voltage is applied the less the capacitance. Use of a PN junction diode in this manner has led to simplification of certain circuits, such as automatic frequency control loops. Here the oscillator frequency is now controlled through varying the capacitance in the circuit by merely applying a potential to a simple junction diode. This is a function that once required a tube and a lot of associated components. It is a clear example of simplification because of a new property of the solid state device. This same property, incidentally, makes possible certain forms of parametric amplifiers. When a signal traverses a circuit comprising a lot of these diodes distributed along a transmission line, the signal itself varies the capacitance, or the capacitance can be varied by some pumping frequency, and by this means parametric amplification can be accomplished.

Now, one would think that this might exhaust the capabilities of this little device, but it does not. We are all familiar with the breakdown diode, or Zener diode, as it is commonly and erroneously called. It conducts current, as all diodes do, in a forward direction but does not conduct in the back direction until a critical voltage is reached. This critical point makes a very good reference voltage for DC voltage comparison. The breakdown voltage may be used to establish the output level in DC power supplies,

or simply to drop potentials in circuits from one value to another and at the same time, transmit variations that may be present. These are things that were a little harder to accomplish before, but that can now be accomplished by the simple junction diode.

Then, of course, there is the tunnel diode, which has its remarkable negative resistance properties. Instead of having the characteristic of an ordinary diode, the tunnel diode conducts current in both the reverse direction and in the forward direction for small voltages, but then as the forward voltage is increased, as shown in Figure I,

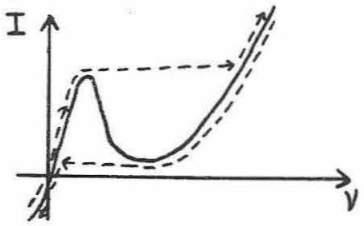


Fig. 1. Voltage-current characteristic of a tunnel diode

the current drops and finally rises again. A high impedance tends to specify the current through the tunnel diode, so that if the current is gradually increased from zero, the operating point moves along the diode characteristic until the peak current is reached. The operating point then shifts rapidly to a higher voltage intersection on the right hand part of the characteristic, and continues to move up this branch as the current is increased. As the current is reduced, the operating point moves back down this branch until the valley current is reached, then shifts rapidly to a lower voltage intersection and continues toward zero. Thus, the unit exhibits hysteresis, and there will be jumps in the output voltage when the input current is slowly varied above the peak and below the valley current.

This behavior resembles that of the so-called Schmitt trigger circuit which employs two vacuum tubes and several components interconnecting these tubes. With a simple resistor and a tunnel diode one can now convert smooth variations, such as sinusoidal variations, into wave forms which are essentially square with very small rise and fall times. This is useful in synchronizing circuits, and in any case where one wants to make an amplitude comparison and generate a sudden pulse when the variable reaches a critical value.

By using a low DC impedance having some series inductance to bias the diode in the negative resistance region, oscillations associated with the inductance and the capacitance of the diode will grow. This permits a very high frequency oscillator to be made very cheaply, and these units will undoubtedly be used for this purpose. In this case a tunnel diode replaces a complete transistor or a complete vacuum tube, and a function we formerly thought required a three-terminal device can now be accomplished with the simple two-terminal diode.

We have been working on a very useful component which around our company was known for a long time as the Boff diode, because it was one of our men, Frank Boff, who first uncovered its unusual characteristics. In some laboratory experiments he was trying to make a harmonic generator which would produce large quantities of a high order harmonic, and the circuit worked better than he thought it had any right to do. We tried to scale this down and look at it on a reduced frequency basis, but then it did not work. We finally resorted to a sampling scope to find out what was happening, and it turned out to be a peculiar effect in the diode itself that was causing the high order of harmonic generation. The effect is as follows: When a diode is forward biased, minority carrier holes are stored in the N-region and minority carrier electrons are stored in the P-region. If the bias is suddenly reversed the stored carriers must diffuse back across the junction before the diode can achieve its normal high reverse resistance. In an ordinary diode, carriers are stored some distance

from the junction, so that upon bias reversal there is an appreciable recovery transient during which the current first reverses to a value limited by the circuit components, and then decreases exponentially as the stragglers of the stored carriers come in from afar.

However, if the doping in the diode is tapered, there is a field built into it which concentrates the stored carriers near the junction. They are then readily accessible again when the bias is reversed. The device then conducts as does an ordinary diode until it runs out of carriers, but it runs out suddenly and the current stops. The door slams shut; there is no more current. This action takes place in a fraction of a nanosecond. There thus results a very fast change in voltage across the diode. This voltage variation can be differentiated to generate short pulses. We now use the Boff diode to produce nanosecond pulses of 8-10 volts in 50 ohm circuits. It is an excellent device for the generation of short pulses.

All of these devices are just PN junctions that have been specially designed to enhance one of the many peculiar properties of this junction. Each one is interesting in itself, but taken together they accomplish so many functions that one might think they leave little need for transistors.

Another solid state device that has found many applications in instruments is the photoconductive cell which we heard described and discussed here previously. We at Hewlett-Packard have used small photoconductive cells about a quarter of an inch in diameter as a replacement for the mechanical chopper in DC amplifiers. One of the cells is connected between the input of the amplifier and the source, the other is placed between the input of the amplifier and ground. Upon alternately illuminating these cells, the input is connected alternately to the source and to the ground. If this is done at a rapid rate, the function is that of a normal mechanical relay chopper. These devices have a very low photovolteic effect and they are very stable with time. They allow very stable, quiet DC amplifiers with no moving parts.

We have also used photoconductive cells for coding and translation operations. In an electronic counter, the information used to be displayed in the form of ten neon lights per decade. Recently designed numerical indicator tubes provide numerals that are larger and more visible at a distance. We now use these so-called nixie tubes together with a translator that goes in front of the original neon lights. The translator simply consists of a common bus bar with ten photoconductive cells tied to it. The other cell terminals go to the various numerals on the nixie tube. When a neon is lit a particular cell conducts and the proper number in the nixie tube lights up. The whole translator is a very compact unit that mounts in front of the original existing equipment.

But photocells can do even more sophisiticated things. On our new transistor counters, for example, they are used to convert from binary logic into decimal read-out, say with neon indicating lights. It takes only a few photocells to do this and at the same time provide storage so that when the unit is counting, the display remains on. A fairly sophisticated operation can be done with a little codeplate, a lot of printed interconnecting wiring, and a few photoconductive cells; a small, quiet package - again, no moving parts.

In connection with photocells, the electroluminescent junction is rather a new development. Again it is a PN junction, but it has an interesting property that all PN junctions have to some extent, but which in this case, has been enhanced. When current flows in the forward direction, minority carriers are produced on the opposite sides of the junction. Under steady state conditions these carriers recombine as fast as they are introduced with proper activation in a proper kind of material (for example, gallium

phosphide). This recombination produces light. So, such a diode is the reverse of the photodiode; it produces light in response to current. It is ideally suited for use with transistors since the impedance, voltage, and current levels are right. Now, if we oppose a photodiode and an electroluminescent diode, we have a true optical relay with no moving parts. It is a true relay in the sense that the controlling circuit and the controlled circuit are electrically separate. They are not connected together electrically and thus do not have to be at a common potential. We circuit designers in electronics have become so used to this restriction that we cannot imagine the freedom that exists for the relay circuit man. It gives a completely new freedom, because there is no longer need to worry about potential differences between the controlling circuit and the controlled circuit. Also, a lot of circuits can be controlled by one controlling circuit. So it is a real, true analog of the mechanical relay and, I think, will probably see considerable exploitation some time in the next few years.

I probably should say a few things about masers and lasers. Actually, we have not made much use of maser amplifiers in any instrumentation yet. The maser as a low-noise amplifier deserves to be included as an integral part of any low-noise, high-performance, expensive receiver, where power is at a premium or circuit loss is high. There seems, however, to be little application for the low-noise amplifier separate from its integral incorporation in a system.

The maser principle is being used in instrumentation in the form of the atomic beam and gas cell devices which achieve a very high order of frequency stability for precise frequency standards. Stabilities of one part in 10^{10} have now become possible because of the use of induced emission in gases for stabilization purposes. In other words, I would say that masers are already used in instrumentation more as sources than as detectors.

The laser is so new that the instrument applications for this have not yet developed, but I am sure they will. Let us look at some things that we might do with the laser.

The outstanding feature of the laser is that it puts out a plane parallel beam of coherent electromagnetic radiation. Such a beam can be focused into a tiny region in space to produce extremely high energy densities. Let us see how high these might be.

You heard yesterday a figure of 10 kw. as the peak pulse output of a typical laser. When a collimated beam of light is focused with a lens, a diffraction pattern of the circular opening of this lens is formed, and if the f number of the lens is fairly high, it turns out that most of the energy falls in an area which is about one square wave length in area. So that if one is talking about a wave length of one micron there is the rather fantastic energy density of 10^{16} watts per square meter at the focus. As a matter of fact, the surface of the sun puts out 64 megawatts per square meter - roughly 10^{7} - 10^{8} watts per square meter - so that with a laser we can achieve an energy density greater than that of the surface of the sun by a factor of 10^{8}.

I do not know how this can be used, but it certainly is an astounding energy density. One ought to be able to do some rather fine molecular stitching with it. It seems likely that electric field intensities would be so high that electrons and atoms would be ripped out of bonds simply by the fields involved. In fact there is some possibility that if two of these laser beams are crossed it might be possible to observe photon-photon interaction. This would be a very interesting physical experiment. Perhaps these intensities can be used for engraving or for other more mundane applications. The diffraction pattern at the focus is as near a point source as can be obtained. In other words, it is the classical point source of optics. If the energy in this "point" source is projected in a beam by a further lens or mirror system, the beam spread that results is determined by the diffraction pattern of the aperture as an antenna,

which is thousands of wave lengths across, rather than by the size of the source. The source is essentially of zero size, and the only beam spread comes from the wave nature of light.

There are many very sophisticated optical instruments that work on the wave nature of light by phase coherence or phase contrast, and the existence of a true point source will make them much easier to design. For example, if one puts a photograph in the plane of a lens imaging a coherent source, what appears in the focal plane is no longer theAiry disc diffraction pattern, i.e., the transform of this circular aperture, but is in fact the transform of the entire picture. It is a complete two-dimensional Fourier transform. The pictorial information is displayed in wave number space with the DC component on axis and the field components appearing off axis by an amount proportional to their wave number. If one wants to attenuate the low frequency field components with respect to the high, a mask that is dense in the center and less dense at the edges can be inserted at this plane. Thus one can do equalization optically rather than electrically. Images of the photograph formed by a lens following the mask can be sharpened or differentiated versions of the object, or indeed, can be the result of some more general operation such as comb filtering.

Another application of the laser that excites the imagination is in the field of very long-distance communication. Let us consider the standard transmission formula, $\frac{Pr}{Pt} = \frac{At\,Ar}{\lambda^2 d^2}$, which gives the ratio of the energy received by a receiver (Pr), to the energy that was transmitted (Pt). At is the area of the transmitting antenna, Ar the area of the receiving antenna, d the distance between them, and λ the wave length. The interest in lasers for long-distance communication stems from the fact that λ is in the denominator and is very small for light. In other words, in going from radio waves, even microwaves, to light, the wave length has gone down by a tremendous factor and so the loss is greatly reduced. Consider X band of 3 cm wave length compared with light of 0.75×10^{-4} cm. The ratio is 4×10^4. However, the payoff is not as great as is indicated by the formula when we increase the frequency, because the noise also increases. If one plots noise power vs. frequency, thermal noise falls off nearly exponentially as we go up in frequency, according to the relation $\frac{h\nu}{e^{\frac{h\nu}{kt}}-1}$. However, this is not the whole story. A thing called spontaneous emission is also present and adds another term $h\nu$. The sum of these two terms gives the total noise power density $N = \frac{h\nu}{e^{\frac{h\nu}{kt}}-1} + h\nu$ which increases monotonically with frequency, ν. Eventually, the noise power density increases proportionally to the operating frequency.

It is interesting to put some figures in the standard transmission formula. Let us assume that At, the area of the transmitting antenna, and Ar, the area of the receiving antenna, which in this case are the apertures of our optical system, are each 20 square meters, the area of the 200-inch telescope. Let us say that the wave length is 0.7 micron, which is a red light, and the distance that we are trying to transmit over is 10^{17} meters, or 10 light years. It then turns out that the ratio of received power to transmitted power is 8×10^{-20}. Now assume a 1 megawatt pulse with a duration of a millisecond, which is on the order of magnitude of the present pulse length. Then the transmitted pulse energy is 10^3 joules, so the received energy now will be 8×10^{-17} joules per pulse. At this frequency one quantum has 3×10^{-19} joules. Consequently, the expectation per received pulses is 300 quanta. With a laser, operated as a coherent detector so that its fluctuation noise was on the order of one quantum, the signal-to-noise ratio would be 300:1 in power, or 25 db, which is not bad.

Robert Heinlein once wrote a science fiction story called "Universe". This was the story of life on an interstellar spaceship, many generations after a mutiny had

occurred which destroyed the original crew. This had happened so long ago that the original reason for the voyage and, indeed the very fact that they were on a voyage, was lost to the memory of everybody on board. They did not know that they were on a spaceship. As a matter of fact, to its human cargo this spaceship was the universe. But the intricate machinery of this spaceship went on functioning perfectly through the millennia because, as Heinlein put it, its builders had builded well indeed. There were no moving parts. All of the activity was at a molecular level as in a transformer. When I look at the level at which some of these complicated devices now are operating, and the level at which some of the complicated electronic functions are being performed by these modern solid state devices, I really feel we are approaching the sort of sophistication that Heinlein described.

The reliability of solid state devices with their cool, molecular level of operation becomes tremendously important in the instrumentation systems field that I mentioned earlier, where as many as several hundred to a thousand automatic data channels may be working simultaneously, gathering all the information to be had from one enormously expensive experiment - or test, such as nuclear explosion or a missile firing. Failure of one or more of these channels could be quite costly. So the trend in these areas is and will continue to be in the direction of completely solid state devices. And, of course, the compactness and low power consumption which we have mentioned pays off in these systems in the form of reducing building space and cooling capacity required for them. So there are some architectural savings as well.

As time goes on and the reliability of the electronic equipment further improves through solid state design, electronic instrumentation will be applied more fully into areas where enormous investments, or where even life itself, may be at stake. I am thinking here of completely automatic process control in large chemical manufacturing plants. Already portions of such processes are controlled automatically, but in general the electronic equipment is not yet reliable and smart enough to be entrusted with the entire plant. Already tape programmed automatic milling machines and tools are turning out individually machined parts of considerable complexity. We are not quite ready to interconnect these machines to form an automatic factory. However, the day is probably not far off when this will happen. We may all live to see it.

When the reliability of an electronic device exceeds that of the human being, for its particular function, the device can be considered and, I think, will be used for applications where today we employ people. Fully automatic aircraft control would be wonderful if, and only if, the enormously complicated systems required were more reliable than human beings. Solid state devices may well make this possible.

Electronic sensing instruments are starting to be used in medicine. With better reliability, they can and will be used and trusted more widely. Banks of sensing instruments, measuring respiration, heart rate, blood pressure, blood sugar, nerve activity, etc., can be directly coupled to computers for diagnostic purposes, or replace anesthetists during operations. The computer can control artificial lungs, a heart, or kidneys, to assume one or more of the patient's organic functions during the operation - during surgery on these organs, for example. This may some day come, but not with tubes. With life at stake one cannot risk having a gassy 12AX7 among a thousand tubes.

There are still other areas where electronics will be applied when the first cost can be justified by the long life of the equipment - such things as teaching machines in the schools, automatic order handling in industry, or an automatic fire and police alarm for the home, which dials help over the telephone when distress comes.

My point is only that these things are possible, given low enough cost and high enough reliability; they are impractical otherwise. It is too early to say, I am sure, whether

man will ever reach the stars. Perhaps the spaceship which Heinlein visualized will forever remain a dream. But it does begin to appear that we may soon be able to communicate with other solar systems by means of the laser, or by some such comparable device. Should this come to pass, it will, of course, revolutionize human thought. But whether it happens or not, solid state devices in thousands of mundane applications will extend the influence of electronics over our lives in the years to come. I am not sure that this is progress, but it will probably happen.